Behavioral Economics & Machine Learning

Expanding the Field Through a New Lens

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2021

vorgelegt
von

Marcel H. Schubert

aus

Aalen (Deutschland)

Referent: Prof. Dr. Martin Fochmann
Koreferent: Prof. Dr. Dr. h.c. Christoph Engel

Tag der Promotion:

Acknowledgments

I gratefully thank my partner, Julia Müller for her unending backing of my many academic pursuits and for always giving me the additional push during these last years – I certainly did not make it easy. I also want to especially thank my parents for their unwavering support, ideationally and materially, thus enabling me to reach my goal.
Moreover, I thank my supervisor Prof. Dr. Dr. hc. Christoph Engel, for being always open to explore a new method or avenue while at the the same time offering keen insights and helpful advice – and generally for being a great research partner and for pointing me in the right direction. Further, I want to thank my supervisor Prof. Dr. Martin Fochmann for keeping me on track and for making sure that I keep the end-goal in sight. It would also be remiss of me not to mention how helpful the comments of Felix Albrecht have been – on any topic really – and thus, I also want to thank him for always having an open ear and some good advice. Additionally, I want to thank all members of the Max Planck Institute for Research on Collective Goods in general. Of these, I am especially grateful to the members of the Ratio Round for always and without fail pointing out any and all potential shortfalls of a research project at any point of its life cycle (and for doing so without sparing me false consideration).
Finally, I thank Dr. Brian Cooper for helping me to make the parts of this dissertation legible for which I did the majority of the writing – he certainly is the one the readers should be most thankful to.

Behavioral Economics & Machine Learning

Expanding the Field Through a New Lens

Marcel H. Schubert

schubert@wiso.uni-koeln.de

Abstract

In this thesis, I investigate central questions in behavioral economics as well as law and economics. I examine well-studied problems through a new methodological lens. The aim is to generate new insights and thus point behavioral scientists to novel analytical tools. To this end, I show how machine learning may be used to build new theories by reducing complexity in experimental economic data. Moreover, I use natural language processing to show how supervised learning can enable the scientific community to expand limited datasets. I also investigate the normative impact of the use of such tools in social science research or decision-making as well as their deficiencies.

Keywords: Behavioral Economics, Experimental Economics, Law and Economics, Empirical Law, Machine Learning, Natural Language Processing, Language and Behavior

Contents

List of Figures

List of Tables

CHAPTER 0

INTRODUCTION

Human behavior is highly varied, even at the individual level. When small-group interactions are taken into account, the variations in observed behavior become complex, so much so that their analysis becomes difficult when only employing conventional tools. Consequently, the discipline of behavioral economics turned to controlled field experiments and lab experiments. Their use has been highly successful and a is a cornerstone of modern behavioral economics. Experiments generate data about specific behavioral phenomena in controlled environments. Thus, they enable researchers to conduct straightforward analyses and to answer precise questions. However, experimental methods are not without limitations. The question of external validity is often raised in the literature (see, e.g., Levitt and List, 2007; Ariely and Norton, 2007; L. Pritchett and Sandefur, 2014). The question, at its core, is how much insight one can gain about the intricacies of human behavior when one (over)simplifies the real world to operationalize the experiment while drawing on a relatively limited sample. At the same time, the rise of mass social media has caused the amount and complexity of observable real-world data to increase tremendously. Many behavioral scientists have sought to capitalize on the availability of such data (see, e.g., Barberá and Rivero, 2015; Flekova et al., 2016) or to expand the scale of their experiments online to approximate the real world more closely, at least in terms of sample size (Horton et al., 2011). However, compared to classic experimental data, the data that researchers are starting to examine now is considerably more complex. That necessitates the adoption of novel analytical methods, not least because the data is commonly textual in form, as is the case with social media. As such, the discipline is being confronted by and willing to engage with these different data types, such as text, whose treatment was fraught with difficulty in the past. Moreover, not only for this new data and not only the analysis of open-form answers entails difficult and tedious manual as well as analytical labor. That is also the case when one seeks to detect patterns within games of punishments and contributions, it

holds even for data generated by traditional lab experiments. Moreover, it often involves coding rules. Those rules, too, were manually generated and therefore might introduce subjectivity and possibly even experimenter bias. Hence, the discipline is at a crossroads as such data precipitate departures from conventional statistical and econometric analysis.

The neighboring discipline of machine learning has developed an extensive arsenal of tools to not only tackle complex and patterned data, but also tools allowing the identification of patterns that previously escaped statistical analysis. Moreover, the cost of deploying such tools and models, in terms of know-how and computation, has decreased considerably (Narayanan et al., 2012). That is why Varian (2014) and S.-H. Chen et al. (2017) have written on the advantages and prospects of such methods, which include reducing the labor intensity of analysis and coding, combating the overspecification problem, and reducing subjectivity in data partitioning and rule specification. Despite their significant theoretical potential, these methods only see sporadic use (e.g., Gerlach et al., 2018; Hausladen et al., 2020). This dissertation is an attempt to bridge the gap. I focus on specific and challenging questions in the behavioral sciences while using new tools that have seen little use in the extant literature. Since policy questions are always latent in behavioral research (Congdon and Shankar, 2015), I also examine the normative implications of using such models. In this way, I hope to not only answer questions on human behavior but also to facilitate the re-examination of old problems and challenging data in the discipline.

To do so, Chapter 1 focuses on the influence of language on decisions and beliefs that pertain to risk and time. Moreover, I tackle the methodological issue of testing their effects when a between-language design is not sufficient. In Chapter 2, I focus on another central field of behavioral economics, behavioral types, by focusing on data from multi-round interaction games, namely public-goods games (PGG). I apply machine learning, in the form of clustering, to such data. The purpose of the exercise is to structure the data and thus try to infer meaningful behavioral types that extend the literature. For Chapter 3, I venture beyond experimental economics, and I study data on judicial decisions. I map them onto an ideological scale, and, crucially, I present a method for extending the limited hand-coded dataset through the use of natural language processing (NLP).
In Chapter 4, I switch sides. Instead of using machine learning to analyze data, I analyze an algorithm that is already in use, and I try to infer how its design affects prediction outcomes and further analyze its normative implications. Finally, in Chapter 5, I focus on the inputs to machine learning algorithms. The questions that I ask concern reliability and stability, both in terms of patterns in the input data and the resultant predictions. Both are central aspects of traditional econometric analysis and are thus relevant to social scientists who wish to harness the power of pretrained machine learning algorithms.

Expanding the Toolbox The dissertation is intended to contribute to various fields of the empirical behavioral sciences and traditional behavioral economics. However, as

mentioned, I also seek to make a substantial methodological contribution to the field of behavioral sciences. As machine learning tools have reshaped many research fields in a manner that would have hitherto been inconceivable, it is not only natural but also essential to expand social scientific methodology, especially in behavioral economics, in a similar manner (Varian, 2014). I try to show how the application of different machine learning methods yields new insights about old problems. Chapter 2 demonstrates the application of unsupervised learning to data from one of the most researched games in behavioral economics, the PGG. In Chapter 3, I leverage supervised learning to expand a small database in a cost-effective manner, increasing its usefulness in research. In contrast, Chapter 4 and Chapter 5 cover the flip-side and look at the normative implications of the use of machine learning models. I also illustrate the limitations of their application to the behavioral sciences. Consequently, this contribution lies at the intersection of traditional behavioral and empirical social science and the technical field from which the methods originate.

Therefore, this dissertation makes a dual contribution to the field. That contribution is substantive, in that new insights are provided about challenging questions in the behavioral sciences, and methodological, in that I attempt to expand the toolbox of the discipline. My approach combines the traditional methods of behavioral economics and the opportunities that emerge from the development of machine learning. It is also my desire to show that the use of these new tools in the social sciences must be considered carefully: they have troubling implications, both normatively and otherwise. My dissertation is thus highly interdisciplinary, bridging as it does the gap between machine learning research and the behavioral sciences.

Individual Contributions to the Literature Both text and language matter in traditional behavioral economics. More often than not, however, language and culture are used interchangeably to control for individuals' characteristics in econometric analysis. In Chapter 1, we focus on this issue and how to distinguish between culture and language within a traditional behavioral setup. K. Chen (2013) was one of the first behavioral economists to argue that language and, more specifically, its grammatical structure, may influence decisions that pertain to the future, independently of cultural background. In brief, his findings are that speakers of languages which require explicit grammatical references to the future, such as English, save less than speakers of languages which do not require the use of such markers. He formulates this as the "linguistic-savings hypothesis". The only deficiency of his findings is that he used observational data. Consequently, claiming causality or even direction is difficult. The hypothesis is not without experimental support (Sutter, Angerer, et al., 2015). The main challenge, however, is this: how does one adequately test for the influence of variation in grammar while holding culture constant?

Using bilingual subjects may sound appealing. It has been shown, though, that switching languages may switch cultural mindsets (Li, 2017), that issue can, however, be circumvented. Not all languages are created equal. There are different language families. The

languages that belong to them follow different rules. After careful consideration, we conduct an experiment in German. In German, future events may be described in the present as well as in the future tense. Consequently, we can deploy a simple between-subject design with grammatical framing as the experimental manipulation. Moreover, we differentiate between decisions and beliefs to test whether one may be affected while the other remains constant. We also include an experimental task to extend our results to the domain of risk, which is said to be interlinked with time in the literature. That combination allows us to test most of the behavioral channels that are liable to be affected by the treatment in a comprehensive manner.
We use the subject pool of the BonnEcon lab but conduct the experiment online using oTree. For the latter, our motivation is twofold. First, we expect the effect to be small in size, and we therefore seek a large number of subjects. Second, we favor an environment that the subjects would find natural in order to not influence any outcomes. We find that choices are not affected by grammatical framing in any way. This holds both for decisions that pertain to time, such as discounting, and for decisions that pertain to risk. As far as beliefs are concerned, we find some evidence of the existence of an effect. However, when we controlling for individual characteristics and other secondary effects, the results lose their significance. We also manage to disrupt the framing effect easily by introducing pre-formulated beliefs. We therefore conclude that there is little evidence of the existence of a behavioral channel that runs from grammatical framing to choices and beliefs.

In the same vain, we go on to examine another core issue in behavioral economics. Chapter 2 focuses on patterned heterogeneity that is present in multi-round interaction games. In terms of economic games, we focus on PGGs. As noted earlier, these datasets are challenging for traditional statistical analysis because the participants interact with each other. Consequently, their choices are not independent, and assumptions that are vital for common models, such as regressions, are violated. Moreover, while it is generally assumed that there are behavioral programs and types, choices made by one individual may also depend on the choices of other members of their group. Consequently, the choices made by the same type may be completely different when the environment changes.
An illustration might be helpful. In the PGG literature, there are five theorized behavioral programs: altruists, conditional cooperators, far-sighted free-riders, hump-shaped contributors, and short-sighted free-riders. Given those five types and a group size of four, there are 35 possible combinations for three players. That means that a fourth player would be exposed to 35 possible environments. If that player is sensitive to the choices of the others, her choices may exhibit 35 different patterns despite her type being constant. This example does not even account for random errors or other outside influences, which may make the resultant patterns noisy. Moreover, analyzing such a rich space demands an amount of data that traditional economic experiments cannot generate, and the manual specification of the type space becomes somewhat difficult.
For these reasons, we employ a purely data-driven approach. We leverage unsupervised

learning in the form of multivariate time-series clustering to partition the data and thus the type space into manageable parts. This approach enables us to refrain from subjective partitioning and eyeballing. Moreover, the number of partitions is not limited by what we expected to find.
One might think that clustering algorithms identify all types directly upon the employment of this method. However, since the clustering algorithm is fed a combination of observed individual choices as well as the observed average contributions of the other group members, what the algorithm truly clusters is similar exhibited behaviors. Consequently, we adopt a two-step procedure. In the first step, we simulate a large dataset that contains all combinations, using the five previously theorized types. We derive two things from that simulated dataset: a data-driven measure for selecting the number of expected clusters, which relies on a combination of cluster validation indices and the variance of the analyzed data and the specification of the cluster algorithm's hyperparameters. As a result, many of the clusters that we find when applying our measure to the simulated dataset are mostly pure, that is, all individuals in a cluster are of the same type. At the same time, some of the clusters are impure because certain groupings produce nearly identical observed choices for different types. For example, a conditional cooperator could be grouped with three free-riders. The conditional cooperator exhibits the same contribution pattern as a hump-shaped player exposed to the same group, and the two may be indistinguishable. We do not view this as a limitation—it shows that the main goal of any data-driven approach is to find distinct pattern types. In the second step, the researcher discusses and tests whether those patterns are generated by a specific behavioral program.
We also validate our approach by using a dataset that consists of 16,474 observations of PGGs from the literature. We show that the true type-space is far richer than theorized. While we find the theorized types, we also isolate previously unexplained behaviors, such as backward logic and strategic, qualified, and cognitive behavioral programs. Finally, we identify two means of narrowing down the set of behavioral patterns found. One is the classic experimental approach. The other is to continue with a machine learning tool by employing symbolic regression to consolidate the clusters further. Our main contributions are to two topics in the field. First, we show the limitations of the theorized type space in PGGs and the means to expand it cost effectively and in an unbiased way. Second, we show how data-driven approaches that rely on unsupervised machine learning may be used to generate new and valuable insights on much-researched problems in behavioral economics.

Similarly, Chapter 3 leverages new machine learning methods in order to contribute to a widely researched topic. The chapter focuses on empirical legal studies and the impact of judicial ideology on adjudicative decision-making. Whether judges lean towards liberalism or conservatism is of considerable interest not only to the legal community but to society as a whole – optimal judicial decisions would be impartial and rest solely on the law. However, there is ample research showing that ideological biases affect the outcomes of adjudications, especially on higher courts (e.g., Jeffrey A Segal et al., 1995; Epstein, Andrew D Martin, et al., 2007; Fischman and Law, 2009; Frank and Bix, 2017). As

such, the aim to ascertain by how much decisions are influenced by such pre-formulated leanings is of considerable interest to the behavioral sciences as well.
In that context, for the first part of the paper we replicate the findings of Landes and Posner (2009), who connected decisions to the political party of the president who had appointed the respective judges. We also expand the analysis by including multiple robustness checks, such as aggregating the dependent variable on a higher level and multiway error clustering. Most of the research on the topic, including that by Landes and Posner (2009) relies on the Songer database (Songer, 1993). That database provides hand-coded labels for approximately 20,000 U.S. appellate decisions. That dataset may appear large, but the total number of U.S. appellate decisions exceeds 1 million at present – less than 5% of all decisions have been coded. Moreover, the decisions in the dataset are not drawn at random from all appellate courts. Instead, the dataset reflects a stratified draw in which each court is represented with the same fraction of decisions. The database is thus substantially limited, and hand-coding even a substantial fraction of the remaining 95% of opinions is too costly to contemplate. In order to overcome this limitation, we employ supervised learning and train a classifier to label the remaining 95% robustly. We achieve this by adopting a tf-idf weighted bag-of-words approach with a calibrated ridge classifier. We then assess the quality and robustness of the classifier by repeating the expanded analysis that is based on Landes and Posner (2009) research with the labels that we have generated. We show that our initial replication holds with the expanded dataset. Therefore, we provide empirical legal scholars with a valuable and cost-effective tool for expanding the amount of data that can be used in research.

In Chapter 4 we assess the predictions and the underlying drivers of an algorithm that is already in use. In this way, we contribute not only to the field of empirical legal studies but also to a wider normative debate in society. The algorithm in question is COMPAS, which is developed and distributed by the private company Northpointe. The algorithm is used in the judicial and penal system of several U.S. states. It is designed to generate a risk score, which is a prediction of the likelihood that an apprehended individual will recidivate. That score is then considered by judges in bail hearings, at sentencing, and when setting probation. Such algorithms and scores are by some seen as a solution to the aforementioned bias-problem present when using human judges (e.g., Danner et al., 2015; Harris and Paul, 2017). However, we show that, here too, normative biases in the form of design decisions may majorly influence the result of such scoring systems. And despite the obvious impact of these scores on the lives of defendants, they are often not entitled to information on its calculation. The U.S. courts have ruled that the internal design of the COMPAS algorithm is the intellectual property of Northpointe and thus protected. ProPublica was able to use the Freedom of Information Act to secure access to scores that the COMPAS algorithm had generated as well as a substantial amount of input variables for 5,759 individuals (Angwin et al., 2013). In the following years, the dataset was used by many of those who wish to ascertain whether the algorithm is racially biased (e.g., Angwin et al., 2013; Fass et al., 2008) or not (e.g., Brennan et al., 2009; Flores et al., 2016).

However, we argue that this debate, which circles on outcomes, obscures deeper issues. That an outcome is biased in one direction or another might be a byproduct of design choices that are more fundamental and perhaps even intentional. In order to substantiate this argument, we first analyze the assumptions that could drive the design of the risk score. The score itself takes values between 1 and 10, where 1 corresponds to "least likely to recidivate" and 10 to "most likely to recidivate." However, these scores are not the output of the model. Instead, the raw output is normed, and individuals are placed into deciles. We show that the predictions process and the process of constructing the bin width of the deciles has a substantial impact on the recidivism risk scores that judges observe. By superimposing a logistic transformation onto the raw scores, we also show that even a high decile score of 9 would only correspond to a recidivism risk of 59.1%. In the second step, we train a correction model on hypothetical judicial decisions. We assume judges decide on the basis of a cutoff for the decile score. For that score and every higher score, the judge decides in favor of jailing. For every score below that cutoff, the judge decides against jailing and in favor of bail. In this way, we mimic the binary outcomes that occur in the real world. We compare those outcomes to the ground truth, that is, to whether the individuals in questions offended within the next two years. We show that significant improvements can be made on the original COMPAS outcomes if one aims to reduce the number of wrongly jailed suspects. That, however, comes at the slight cost of failing to incarcerate some true recidivists, that is, at the cost of more false negatives. We also show that our correction improves the situation of disproportionately affected minorities, such as young individuals or blacks. We do not argue that our model is better—our purpose is merely to show that a different outcome on the Pareto front is achievable. Consequently, the outcomes that COMPAS produces may be driven by intentional design choices, such as reducing the number of false negatives at the expense of other errors. That minorities are adversely affected is most likely a byproduct of these decisions. Therefore, we argue that the normative decisions that are buried in the design of COMPAS ought to be made transparent. Otherwise, legislators and judges cannot make decisions in accordance with their normative convictions.

Finally, Chapter 5, presents an evaluation of the impact of small variations in input on the internal stability of an NLP model. On first impression, the issue may seem technical and peripheral to the pursuits of behavioral and social scientists. However, the possibility of using machine learning models, especially ones that can work with textual data, and applying them to large online datasets scraped from platforms such as Twitter excites considerable interest among social scientists and lawyers as well as law enforcement. On one hand, for empiricists, these datasets were previously too large to work with and, more crucially, often lacked essential information about personal characteristics. However, NLP author profiling models can estimate missing characteristics, which may then be used for further research. On the other hand, law enforcement officials apply tools of this kind to gain information about anonymous online offenders. Likewise, decision-makers cite results from NLP profiling tools during criminal proceedings and judicial proceedings. Therefore, it is crucial to understand how robust these tools are to slight variations in

the underlying dataset. However, little has been done in that regard so far (Neal et al., 2017; Rocha et al., 2017).
To tackle that I conduct the following analysis: First, I construct a Support Vector Machine (SVM) model, as suggested in the authorship analysis literature. It aims to predict outcomes from textual features alone, with little contextual information about topics or the domain of the data. I then conduct a controlled machine learning experiment using a precompiled Twitter dataset. I construct sub-datasets that were identical in every aspect except one, the variation in the number of authors that are present in the dataset. I train the same model on each variation, and I try to predict two characteristics that are central to the social sciences, gender and age. While I show that the performance of the classifier remains stable and that it is comparable to what may be found in the literature, I also show that there are specific authors for which the classifier makes systematic errors. For these authors, the accuracy of the results falls short of the random-guess threshold. The implication is that such individuals are systematically disadvantaged when scrutinized by a system that analyzes their individual presence.
In the second step, I examine the internals of the classifier, specifically its weight matrix. The weight matrix holds all the individual coefficients by which an input instance is multiplied to generate the output. Thus, the weights correspond to the directed relevance of individual input features. I analyze the stability of the relevance of the individual features when the number of authors in the dataset changes. For this reason, I calculate their relevance in terms of ranking before introducing the variation and compare it to the new relevance when training the classifier on the modified dataset. I find that overall predictive power remains within the margins identified in the literature, but the predictions for the different sub-datasets are driven by completely different features, that is, their relevance is not stable. When conducting these experiments, I test whether the number of feature types or the length of input texts changes the results. It does not.
It follows from the foregoing that the classifier relies mainly on correlational patterns to predict outcomes and that these patterns are not stable. There appears to be no causal relationship between input and output. It is also troubling that small variations impact the correlations in such a way. Real users would not always train the model. The advantage of ML models lies in the possibility of using pretrained models to predict missing information. However, since the underlying patterns in the data change and some authors are systematically disadvantaged, such models must be applied cautiously. The normative implications are considerable because it is unclear at what point the use-case data becomes too far removed from the training data. As of yet, there are no accepted measures for such, and models are not accompanied with explanations of the underlying boundaries for training vs. use-case dataset. Consequently, the utility of such models for social scientific research is affected. The discipline prides itself on econometric and statistical rigor. However, the application of pretrained models without an assessment of the similarity between the training data and the use-case data might invite undesirable statistical distortions of experimental results and their subsequent evaluation.

Author Contributions. The research on the individual chapters was undertaken with coauthors. Presently, their contributions will be described below.
Chapter 1 was coauthored by Felix Albrecht (FA). The literature gap and the research question were formulated by FA. (MHS) designed the experimental tasks together with FA. The experimental tasks were implemented by MHS and the final implementation was conducted by FA. The econometric analysis was executed by MHS The first draft of the chapter was written by MHS, with FA reviewing the draft and finalizing the text. The research benefited from the financial support of the DFG [Deutsche Forschungsgesellschaft; Grant 50130225]
Chapter 2 was coauthored by Christoph Engel (CE) and Carina I. Hausladen (CIH). CIH implemented the local regression approach that CE proposed and visualized the results. CIH also presented the project at a lab meeting of the Amsterdam Cooperation Lab at VU Amsterdam (February 2020). CIH and MHS were equally involved in data curation and tested different configurations and datasetups. CE and MHS wrote the code for simulating the data. MHS implemented the simulation on the cluster and evaluated the results. CE identified the literature gap, formulated the research goals, and supervised the implementation of the clustering specifications. MHS also implemented and evaluated the symbolic regressions. CIH and MHS drafted the methods section. CE reviewed the latter critically and wrote the remaining parts of the paper, including the sections "Experimental Data" and "Rationalization."
Chapter 3 was coauthored by CIH and Elliott Ash (EA). CIH prepared the data that was used for replication and conducted the regression analysis and the robustness checks. CIH developed and implemented the original version of the code that tested the initial classification setups. MHS was responsible for scraping and preprocessing the text data. Furthermore, MHS implemented the final grid search on the cluster. CIH and MHS were equally involved in writing the original draft and presenting the paper at the PELS Replication Conference in Claremont (April 2019). CIH presented the project at the MPI lab meeting (November 2019). EA supervised the research activity, proposed analysis methods of analysis and visualization, and provided part of the data to be analyzed. He also reviewed the draft extensively.
Chapter 4 was coauthored by CE, Sebastian Lapuschkin (SL), Lorenz Linhardt (LL), and Marina M.-C. Höhne née Vidovic (MMCH). CE identified the literature gap and the research question. MHS preprocessed the data and conducted their initial analysis. LL implemented the first model setup and evaluated it extensively through a parameter search. MHS implemented the distributional analysis of the COMPAS data. LL and MHS contributed equally to the final model implementation and the evaluation. MMCH and SL contributed extensive technical and analytical supervision. MHS and LL wrote a draft of the results section and the methods section, which were reviewed critically by CE. CE also wrote the remaining parts of the paper as well as the final text.
Chapter 5 is authored by MHS alone. MHS thus identified the gap in the literature, and took sole responsibility for implementations and evaluations as well as the composition of the paper.

CHAPTER 1

The Effect of Grammatical Variation on Economic Behavior

Varying Future Time References within the German Language

Abstract: We test the proposed impact of future-tense reference on economic decision-making. To this end, we implement a within language framing experiment, varying exclusively the grammatical reference of future events. We do so by leveraging the grammatical structure of the German language, thereby avoiding the introduction of potential confounds, present in cross-lingual studies. In our results, we find no supporting evidence for a causal link between a language's grammatical structure and the speaker's economic decision-making in the time discounting and risk domain. We find weak support for impacts on individuals' believe formation. Our results hint at the fact that a language or grammar dummy absorbs facets of culture not captured by a culture dummy.

Keywords: Time Preferences; Risk Aversion; Grammatical Framing; Language; Experimental Economics

Funding: This work was financially supported by DFG [Deutsche Forschungsgesellschaft; grant 50130225].

1.1 Introduction

Known as the 'Sapir-Whorf Hypothesis' among linguists (e.g., Regier and Kay, 2009), research has long since conjectured that language-specific idiosyncrasies, like differences in grammatical structures (e.g., Cook et al., 2006; Daniel L. Everett, 2005; Daniel L Everett, 2012), distinct means to describe physical properties (e.g., colors Winawer et al., 2007; Franklin et al., 2008), or non-physical occurrences like emotions (e.g., Lindquist et al., 2006), can affect human behavior and perception.

Cook et al. (2006) show for bilingual English-speaking subjects from Japan scoring highly on English tests, that they exhibit significant tendencies to classify material objects more in line with US than with Japanese monolinguals. Winawer et al. (2007), testing Russian and English native speakers for their color discrimination capabilities, find Russian native speakers to be faster in differentiating colors close in spectrum compared to English native speakers. They attribute the advantage of Russian native speakers to the more diverse color nomenclature available in the Russian language compared to English. Majid et al. (2004) find that frames of references in spatial tasks varied cognitively with the linguistic differences of the respective native languages of children.

Investigating grammatical peculiarities of different languages, the seminal contribution by K. Chen (2013) studies the impact of future-time reference (henceforth FTR) on saving and health-oriented behavior. FTR classifies how strongly descriptions of future and current events are grammatically segregated. Strong-FTR languages mandate the use of specific grammatical indicators when talking about the future, contrary to weak-FTR languages, where future events can be referenced using the present tense.[1] Studying World Bank savings data, K. Chen (2013) finds a strong relationship between weak-FTR languages and higher rates of savings or lower rates of types of behavior detrimental to health. He postulates that two channels may influence his results. Either weak-FTR languages let future events appear more immediate (Linguistic-Savings Hypothesis) or weak-FTR languages cause an imprecision of beliefs about the timing of a future event, also making saving more attractive. However, given potential confounds in the underlying non-experimental real world-data, K. Chen (2013) cautions from interpreting his results as causal. He states that the direction of the linkage of language and behavior is unclear and that language might be a reflection of "deeper differences" transported with the language itself.

To test K. Chen's 2013 language behavior relationship, Sutter, Angerer, et al. (2015) implement a set of delayed gratification task experiments with German (weak-FTR) and Italian (strong-FTR) native tongue school children in southern Switzerland. Sutter, Angerer, et al. (2015) find that German native-tongue schoolchildren show a significant inclination to delay gratification longer, concluding that strong-FTR languages indeed induce higher impatience in their native speakers. Li (2017) uses bilingual (English-Chinese) Hong Kong citizens to test the impact of Chinese and English framing on risk and prosocial behavior. Li (2017) finds suggestive evidence that subjects change their

[1] Examples for *strong-FTR* languages are: English, Arabic, Italian, and Korean. Exemplary for *weak-FTR* languages are German, Japanese, and Brazilian Portuguese.

beliefs about the behavior of others if tasks are presented in language frames differing in FTR. Additionally, Li (2017) finds a preference for "Chinese lucky numbers" when the experiment is framed in Chinese, indicating differing cultural mindsets caused by a change in lingual frame. In our opinion, both results relate to "deeper differences" cautioned by K. Chen (2013).[2] Literal translations in different languages have been shown to evoke diverging concepts and transmit particular supplementary information in the respective language (Houser et al., 2004; Majid et al., 2004; Briley et al., 2005; Luna et al., 2008; Van Nes et al., 2010).[3] A recent study by Thompson et al. (2020), investigating word meanings using semantic alignment across different languages, shows low correlations in all investigated domains, supporting the idea of deviating concepts contained in literal translations. Consequently, a clean identification strategy, aiming at testing whether addressing a future event in the future or present tense impacts economic decision-making, needs to avoid the confounds described above. Ideally the language is kept constant across treatments to avoid the transmission of deviating concepts and or varying cultural mindset. That is what J. Chen et al. (2019) try to test within the Chinese language. To that end, they make use of the fact the Chinese language does not demand a future-reference within every sentence. Instead, the reference may be omitted. However, the drawback is that they do not necessarily test for the effect of the precision of beliefs related to future references. Rather they test for the effect of awareness of future-reference, moving one event closer to the present compared to another. Moreover, they only test for choice tasks and do omit any possible interaction between risk and time (B. J. Andreoni and Sprenger, 2012). We overcome these challenges by leveraging a grammatical feature of the German language. The German language allows identical future events to be referenced in the present and the future tense equally without becoming grammatically false (Dahl, 2000; Dahl and Velupillai, 2005; K. Chen, 2013). According to K. Chen (2013), German is classified as a weak-FTR language as it does not necessitate a grammatical marker when referencing the future. However, the German language still incorporates specific grammatical markers for the future tense. German offers different options to a speaker for referencing future events while staying grammatically correct (Dahl, 2000; Dahl and Velupillai, 2005). One option is to reference future events in the future tense, which necessitates the use of a specific grammatical marker. The second option is to reference future events in the present-tense in conjunction with at least one unspecific temporal marker.[4] The unspecific temporal marker, though not necessary, can equally be integrated in the future reference by future tense. To illustrate Table 1.1 provides an

[2]While the discipline of linguistics does not know the term *belief*, it employs the term *perception* in a similar fashion. Under that terminological umbrella, linguists cluster together those lingual effects affecting *visual* perception (e.g., Athanasopoulos et al., 2010) as well as such effects affecting the *abstract* perceptions of concepts, such as duration (e.g., Bylund and Athanasopoulos, 2017) or professional ideas (e.g., Monti-Belkaoui and Belkaoui, 1983). For this reason, it is not unreasonable to draw the connection between the latter and what economists define under the term *beliefs*.

[3]An example is the concept of "police". While in nations with low corruption and high trust in state authorities, "police" is often connoted with the concept of "helper", this certainly is not the case for people living in states where corruption is rife and the populace is violently suppressed by state authorities. While this is a very salient example, the issue is still valid for smaller and less obvious conceptual differences.

[4]Unspecific temporal markers in this sense are words such as *soon* or *afterwards*.

Table 1.1: Comparison of Future Time Reference Options in English and German

	English	German
Future	I am *going to* buy groceries soon.	Ich *werde* bald Lebensmittel einkaufen gehen.
Present	**Incorrect:** I buy groceries soon.	Ich gehe bald Lebensmittel einkaufen.

- Grammatical future time reference marker is shown in *italic*
- Unspecified temporal marker is underlined

example, comparing English and German language FTR properties. The sample sentence shows that there is only one correct way to refer to a future event within the English language. English necessitates a grammatical marker for future event referencing, in the example "going to", and is classified as strong-FTR by K. Chen (2013). The German language, like English, can express future events using a grammatical marker which shifts the verb to a future tense form. In the following, we refer to this grammatical construct as the future-tense future reference (henceforth FF). However, it is also possible the use the verb in a present tense form if an unspecific temporal marker exists. In the example, this temporal marker is "soon" (German: "bald"). We will refer to this grammatical construct as the present-tense future reference (henceforth PF).[5]

This feature of the German language allows us to vary solely how future events are grammatically referenced and permits us to investigate the impact of FTR on subjects' economic decision-making, while avoiding the introduction of confounds contained in multilingual experiments. Making use of this particularity of the German language, our study tries to provide a clean identification strategy for the effect of future tense reference on economic decision-making within a single language.

In order to investigate the effect of differing FTR on economic decision-making, we implement a delayed-gratification and a risk-aversion task, varying German FF and PF between subjects. To account for potential shifts in beliefs, we further implement a number of vignettes to disentangle whether varying the grammatical frame for future events influences subjects' judgement of the likelihood and immediacy of future events.[6] To our knowledge, we are the first to investigate the causal link between grammatical variations in framing of future events and people's risk and time preferences within a single language.

Comparing subjects' behavior in PF vs. FF framing in the German language, we find little evidence for an impact on people's risk aversion and time preferences. We find weak support for an impact on people's beliefs about the immediacy and probability of events occurring in the future. The impact on beliefs, however, appears to be easily

[5]Note that the unspecified temporal marker is used in both sentences. While it is not required in an FF grammatical structure, as the specific temporal marker indicates a future setting, it is required in a PF structure. As it is required in a PF structure and allowed in an FF structure, we always included the unspecific temporal marker in both framings during the experiment in order to vary only the grammatical tense of the verb.

[6]Framing something in the present tense might convey the event as being more likely and/or more immediate to occur than a framing in grammatical future tense (K. Chen, 2013).

overpowered by preconceived notions held by the subjects. Personal preferences for one or the other grammatical structure do not seem to play a role in subjects' choices.

The paper proceeds as follows. In the following Section 1.2 we outline our approach as well as the experimental design of the study. The subsequent Section 1.3 presents our findings. Section 1.4 summarizes the findings.

1.2 Experimental Design

We implement a between-subjects experimental design to elicit the effect of varying the way future events are grammatically referenced in the German language on the time and risk preferences of German native speakers. We also implement a number of vignettes to investigate potential impact on the subjects' believes about immediacy and likelihood of an event occurring. To this end, we specifically designed the experimental texts to include clean and unobtrusive grammatical variations allowing for a strong framing.

Additionally, we designed a task to elicit subjects' preferences for a specific grammatical tense, which might mitigate the efficacy of the respective framing. Concluding the experiment, subjects answered a socioeconomic survey. With the exception of the task investigating subjects' preferences for a grammatical tense and the socioeconomic survey, every bit of text referencing future occurrences is framed in either PF or FF. This includes task descriptions, vignettes, as well as introductory explanations about the experimental session, behavioral rules, and matters of payment.

Time-Preferences – Choice List

The time-preferences elicitation task, henceforth called TimeGame, is a choice between 10 different payment schemes. [7] The implementation is mathematically equivalent to the traditional choice list approach (e.g., Andersen et al., 2008). Each payment scheme corresponds to a small interval of discount rates. Each scheme could be inspected by subjects before selecting the preferred one. A selection could easily be changed until final submission. Before submitting their choice, at which point no change would be possible anymore, a popup window would ask subjects to confirm their choice. Payment schemes beginning at a later date paid more money overall. The payment schemes are constructed in such a way that they all pay a fixed amount of money once a week over six consecutive weeks. The money is transferred to subjects' bank accounts on the dates corresponding to the chosen payment schemes. All payment schemes start at least one week after the experiment ended to avoid any biases (Burks et al., 2012; Benhabib et al., 2010) or confounds introduced by participating at a date later than a payoff scheme's start. We opted for a simple task because an online experiment only offers limited means to explain an assignment (Dave et al., 2010).[8,9]

[7] See supplementary online materials for screenshots of the experiment

[8] Other tasks for eliciting time preferences were considered, e.g., a nested choice list based on Attema et al. (2015), but deemed unsuitable for an online experiment during a trial runs due to complexity.

[9] section A.2 provides detailed calculations.

Risk-Preferences – The BombGame

We elicited the risk aversion of individuals using a variation of the BombGame introduced in Crosetto and Filippin (2013). During the selection stage of the BombGame, subjects are presented with 100 numbered fields. 99 fields hold cash prices, each valued at € 0.20. The remaining field contains a destructive option, represented by a bomb, which is placed in the 10×10 matrix at random. If selected, the bomb field nullifies all gains from the other fields, leaving the subjects with a zero payoff from this task. [10] Subjects select as many fields as they like, but are not informed of the selected fields' content. The result from this task is revealed on a later page in order to not affect the decision for the other tasks. During the BombGame subjects construct their own preferred lottery. Therefore, as the risk of selecting the bomb field increases with the absolute number of selected fields, this task provides us with a good measure of a subject's risk preferences.

The order of the risk-preference and time-preference elicitation tasks is randomized on the subject level to control for possible order effects.

Belief Elicitation Vignettes

The belief elicitation vignettes consist of small paragraphs in the range of three to five sentences. In total, eight vignettes are shown to subjects, consisting of four vignettes concerning likelihood and four concerning immediacy of events.[11] To the authors' knowledge, no vignettes existed prior to this study to investigate this relationship. Consequently, the vignettes are specifically designed for our experiment. They, too, were tested in a small pilot study with non-incentivised subjects prior to implementation.

Immediacy

The first set of vignettes investigates whether varying FTR influences the perceived immediacy of an event when no explicit information of a future date is provided. Subjects choose the point in time they think reflects the average of what the other subjects estimated to be the most likely time frame when the described event would occur. The time frame could be chosen from ten predefined intervals, ranging from "within a week" to "later than 6 months". Payments depend on a subject's answer either being adjacent to the average prediction of all answers (€ 0.50) or guessing the exact average prediction (€ 1.00).

Likelihood

To investigate the domain of risk, we elicit whether varying FTR influences to the perceived likelihood of events. During a vignette, subjects are asked to guess the average likelihood of an event occurring in the future as indicated by the other participants. Input is possible either via a slider ranging between $[0 : 100]$ or via direct numerical

[10] See screenshot in the supplementary online material.

[11] See section A.3 of the appendix for a complete overview of the vignettes in FF and PF framing.

input. Manually moving the slider sets the numerical value. The slider button is initially hidden until the slider bar is clicked at to prevent setting an anchor. The slider is also color-coded (red for less likely and green for more likely) to improve usability during the online experiment. [12] Subjects would receive a payment of € 0.50 if their answer lies within five percent of the average of all subjects' answers for this vignette, given the same treatment. If subjects managed to guess the correct average, they would receive a payment of € 1.00.

Fischbacher, Gächter, Bardsley, et al. (2010) show that such incentive schemes are a reliable method for the elicitation of beliefs, while at the same time reducing the possibilities for hedging. The likelihood and immediacy vignettes are presented to subjects in blocks of the same type to avoid confusion about what needs to be assessed. The order of the blocks and the order of the vignettes within each block are randomized on the subject level.

As the experiment was conducted as an online study, we designed vignettes in such a way that the content, while easy to understand and believable, had to be completely fictional. This means no factual information about dates in regard to the content could be found. However, while no factual content can be found online, depending on the topic of the vignette, it may be possible either to find related information or come into the experiment with a strong preconception about the topic. The latter could interfere with an induced framing effect. At the same time, we can exploit such an effect to serve as a boundary on the stability of a grammatical framing effect. Consequently, while we had four vignettes in both categories, we opted to include one topic in each category which would offer the possibility to invite additional information, outside preconceptions, or predispositions. The immediacy vignettes contain a scenario incorporating the topic of Bitcoin.[13] The likelihood vignettes include a topic concerning the German city of Buxtehude, which offers itself to preconceived notions.[14] These are the vignettes labelled (d) and (h), respectively.

Concerning the content and topic of the other vignettes, we could not deduce any systematic outside influence, which means that only such information as provided in the respective vignette is available to subjects. The likelihood vignettes covered topics ranging from announcements of a German Federal authority and a multinational consulting firm to a business forecast of a European airport. For the immediacy vignettes, topics touched upon the Bonn EconLab, announcements concerning roadworks, and the scheduling of a city council meeting regarding broadband internet. We chose the design of the vignettes to include specific scenarios as the vignettes needed to be easily understood,

[12] See the supplementary online material section for screenshots of the experiment

[13] The topic was widely discussed at the time of the experiment, exhibiting a high prevalence in newspaper articles and online blog posts.

[14] Buxtehude is a city in northern Germany which many southern Germans consider a proverbial place (Förste, 1995) 'where nobody wants to go or nothing ever happens, i.e., a faraway, place of no concern to anybodyt'. Many are surprised when they learn of its real existence. The real city is located near Hamburg. A potential origin of this peculiar association is that the city prominently featured in a children's tale by Ottfried PreuSSler in the early 1960s. The city would be a faraway place where a warlock would go using his broomstick (PreuSSler, 1962). The story has been continuously retold in popular media and the city features prominently in modern day German children's tales. (e.g., Watson et al., 1992; Bartos-Höppner, 2010; Michael, 2018).

believable, and feature plausible occurrences as the plausibility of the content necessitates a well-considered answer. In a small pilot, no evidence was found that subjects found the content unbelievable or hard to grasp.

Paragraph Construction and Questionnaire

The paragraph construction task elicits subjects' preferences for a tense used to express future events in the German language. Preferences for a specific tense might impact the efficacy of the respective treatment. To this end, subjects have to construct a paragraph consisting of 5 sentences. For each sentence, two versions are provided to each subject; one is in the present tense, while the other is in the future tense. A subject then has to decide which combination of sentences she considers to feel most natural. Since this task potentially draws the subjects' attention to the linguistic aspect of the experiment, the paragraph construction task is placed after all other tasks. As subjects select their preferred tense for a specific sentence, a complete paragraph is generated and shown to subjects, who can make changes to their choices before submission. The paragraph contains a short news report about an urban redevelopment project and future plans for the area. Just as in the vignettes, we chose a topic that was easy to grasp and could not be related to any factual occurrences in a systematic way.

Finally, subjects filled in a survey. The survey includes self-reported measures for risk aversion (Dohmen et al., 2011) and elements from the German SOEP (Wagner et al., 2007), as well as questions on socioeconomic characteristics. [15]

Implementation

The experiments were programmed and conducted as online experiments using the experimental software oTree (Daniel L Chen et al., 2016). This was done to allow for a larger number of participants to pick up on potential null-effects. In order to assess the criteria under which our design is able to pick up on potential null-effects, we primarily considered the paper by K. Chen (2013), as it is the most widely-citet in this domain. The author finds that the odds of an individual saving within the year is twice as high for wFTR speakers compared to sFTR speakers. That effect level remains, regardless of the controls added. While such an effect may conventionally be seen as a strong effect, we used the data available from K. Chen (2013) [16] to estimate the required number of individuals. Consequently, in order to test for an effect size of the magnitude found in K. Chen (2013) with a power in the 95% confidence interval, we require a minimum of 1,083 observations under our experimental design. We therefore opted for the implementation as an online experiment, using subjects from the Bonn EconLab's subject pool.[17] Usage

[15] For a comprehensive list, see the supplementary online material.

[16] The data as well as the code may be downloaded from www.openicpsr.org. We adapted the code and used the R packages pwr and effsize for the calculations.

[17] Online experiments were shown to yield reliable results when compared to traditional lab studies, despite lower stakes (Paolacci et al., 2010; Amir, Rand, et al., 2012).

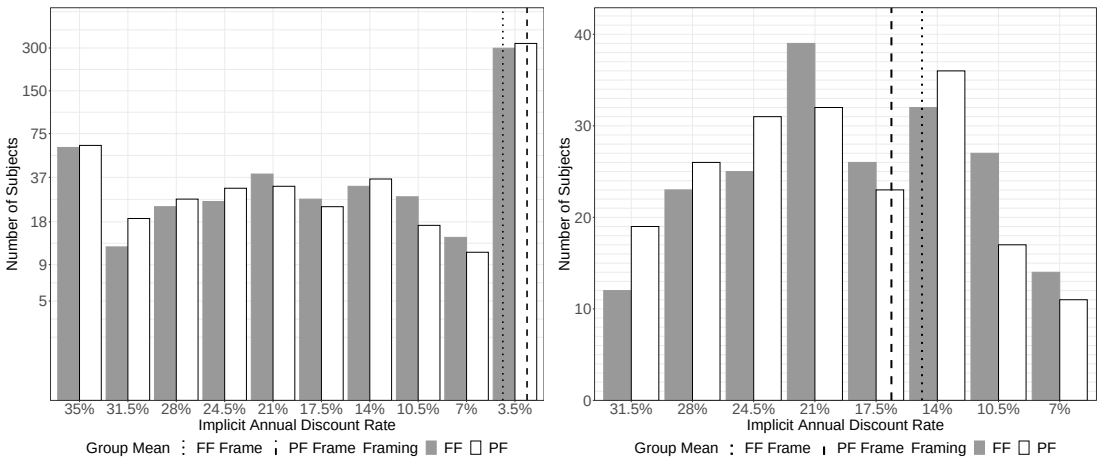
of the experimental software oTree (Daniel L Chen et al., 2016) allowed us to design the experiment to be accessible from a large variety of devices, especially mobile devices.

1,389 subjects participated during the online session and were assigned to one of two treatments – Present Tense Future Reference (PF) or future tense Future Reference (FF) framing – at random. 234 individuals were excluded from the analysis because they either had not completed the experiment in full or had participated twice, in which case only the observations from their first participation are included. Observations of the remaining 1137 subjects are considered in our analyses. Of these subjects, 557 were treated with the future and 580 with the present tense framing.

1.3 Results

Time Preference Elicitation Task and Risk-Aversion Task

Figure 1.1a presents the distribution of decisions in the time-preference elicitation task. A $\tilde{\chi}^2$-test yields a p-value$^{\text{Time}} = 0.674$, leading us not to reject the null-hypothesis of no significant difference in chosen payoff schemes between the two treatments. As the large number of responses at the lower and upper limit could mask existing distributional differences, we remove the observations in the lower and upper margins for a secondary test. The distribution of the reduced dataset is shown in figure 1.1b. A $\tilde{\chi}^2$-test on the reduced set yields a p-value$^{\text{TimeNC}} = 0.525$, showing no significant differences in payoff

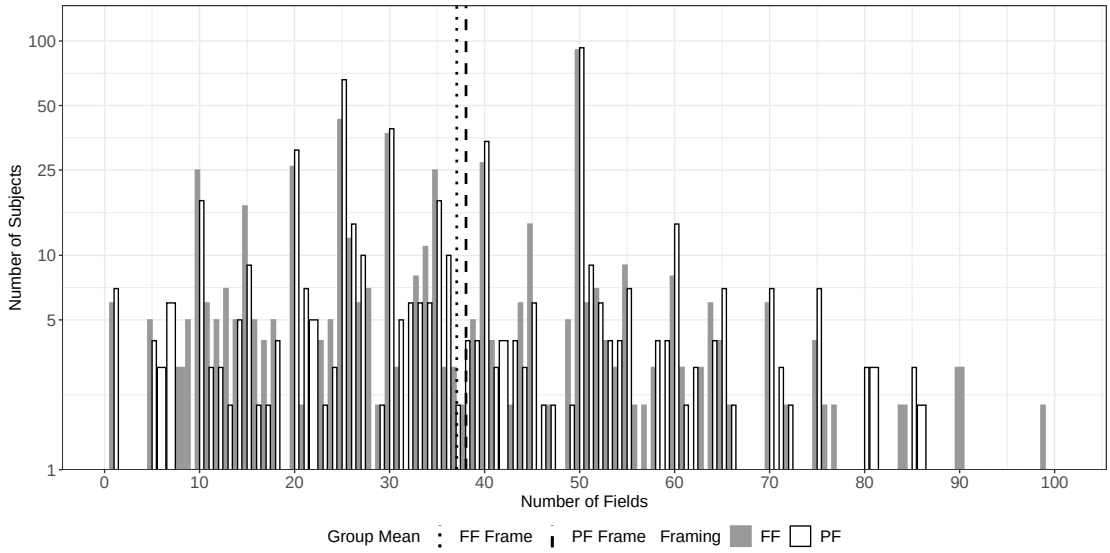


(a) Complete distribution

(b) Lower and upper margins removed

Note: Distribution of choices by treatment for the time-preference task. Panel (a) shows the distribution for 1,137 (557 FF; 580 PF) observations; panel (b) shows the distribution for 393 (198 FF; 195 PF) observations, given that the lower and upper margins are removed. The dashed lines show the median of choices for the respective treatment. Presentation in absolute values.

Figure 1.1: Distribution of TimeGame Choices



Note: Distribution of the number of individually selected lottery fields during the risk-aversion elicitation task for 1,137 (557 FF; 580 PF) subjects. The dashed lines indicate the respective median. Presentation in absolute values.

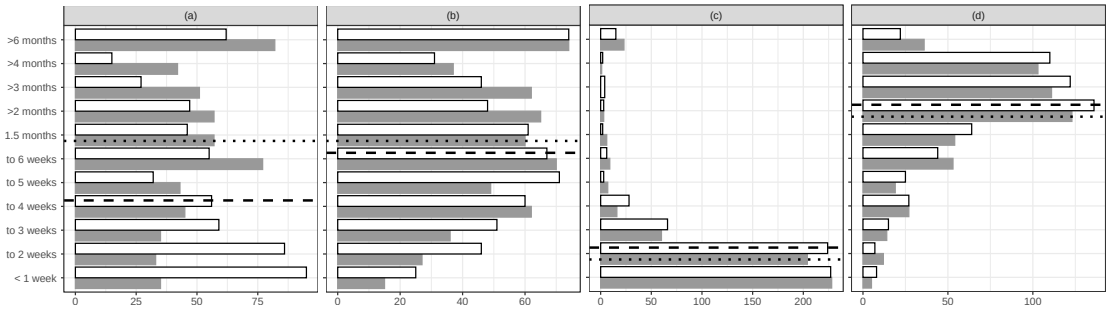
Figure 1.2: Number of Selected Fields in the BombGame

scheme selection between the two treatments. Results from Tobit regressions, presented in section A.2 of the Appendix, support these results.

In the next step, we analyze the results of the risk-aversion elicitation task. Figure 1.2 displays the distributions of individually selected number of lottery fields in each treatment. The visual representation suggests that choices in the risk-aversion task match closely across treatments. This result is supported by a Mann-Whitney-Wilcoxon test (U-test) which yields a p-value$^{\text{Bomb}} = 0.364$. Given the observations in our sample, we cannot reject the null-hypotheses for equality of distributions for the time preference and risk preference elicitation task. The implemented future and present tense framing does not appear to impact the economic choices to statistically significant degrees in our chosen settings. Results from OLS and logit regressions, presented in Table A.4 and Table A.6, support these results.

Belief Vignettes

In the following, we analyze the data obtained from the vignettes. We first focus on the vignettes concerning the immediacy of events. To recap, we implement two types of vignettes. 'Immediacy vignettes' elicit whether different grammatical framings influence subjects' beliefs about the immediacy of future events whose exact occurrence is undetermined. 'Likelihood vignettes' investigate whether the grammatical framing alters subjects' beliefs on the likelihood of events occurring in the future. Vignette (d)



Note: Distribution of subjects' choices for immediacy vignettes. Light color represents FF, dark color represents PF framing. Median of FF is depicted by the dashed, PF by the dotted lines. x-axis presents choices in absolute terms for 557 (FF) and 580(PF) subjects.

Figure 1.3: Distribution of Choices in Immediacy Vignette

and (h) contain the aforementioned checks for the influence of preconceptions.

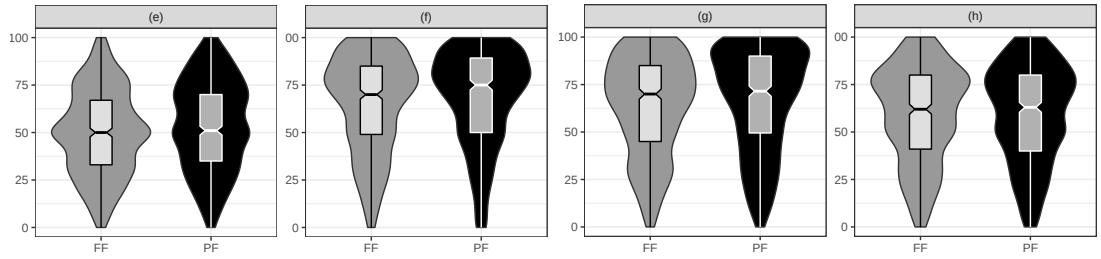
Immediacy

Figure 1.3 depicts the observed distributions of choices made in the four immediacy vignettes. The immediacy vignettes (a) and (b) show significant differences ((a) p-value < 0.01; (b) p-value $= 0.038$; χ^2-test) in subjects' choices about the perceived immediacy of the described events. These results are in line with the idea that future events framed in the present-tense are perceived as more immediate than future events framed in the future tense. The third vignette (c) shows no significant differences in choices (p-value < 0.161; χ^2-test) but suffers from lower-bound truncation. Vignette (d), which we included in the test for the influence of preconceptions, is not significant on any conventional level (p-value $= 0.545$; χ^2-test). Vignette (d) shows median choices for FF which are more immediate than PF choices. That, however, goes against the idea of present tense framed events being perceived as more immediate than future tense framed events. Results for vignette (d) hint at the fact that the PF framing induced believe changes are not robust when introducing additional preconceptions. To check the robustness of these results we estimated ordered logit models, including order effects, the measured preference for a specific tense and socioeconomic data collected during the survey, which support these findings. See Table A.9 in the Appendix.

Likelihood Vignettes

The following section takes a look at the vignettes related to the domain of risk. While the grammatical framing is not directly related to this domain, it may be influenced by an indirect link (Andersen et al., 2008).

Figure 1.4 plots the observed likelihoods for the described events to occur in the future. The plots for vignettes (e), (f), and (g) show a shift in the distributions towards higher



Notes: Violin plots depict the distributions of likelihoods of the described events to occur in the future as assigned by the subjects. Boxplots show the median and 25%, 75% quantiles, respectively. Median: e: FF=50, PF=51; f: FF=70, PF=75; g: FF=70, PF=71; h FF=62, PF=63. Likelihoods could only be chosen using integer values.

Figure 1.4: Violin Boxplots for Likelihood Vignettes

likelihoods in the PF framing, i. e., future events framed in the present-tense are assigned a higher likelihood of occuring in the future. These visual findings are accompanied by shifts in the respective median value (Median: e: FF=50<PF=51; f: FF=70<PF=75; g: FF=70<PF=71.5; h: FF=62<PF=63) and are significant on at least the 5 percent level (p-value$^{(e)}$ = 0.022, p-value$^{(f)}$=0.002, p-value$^{(g)}$=0.033; U-test). For vignette (h), we observe no such clear shift, a finding that is supported by the results of a U-test (p-value$^{(h)}$ = 0.886). Results for vignette (h) similarily to the results for vignette (d), indicate that, although grammatical tense framing impacts beliefs in intuitive ways, the effect is not necessarily robust against additional preconceptions and might be overruled by such. To check the robustness of these results, we estimated ordered logit models for vignettes (e-h), including order effects, the measured preference for a specific tense and socio-economic data collected during the survey. These estimates, although still showing the expected direction of effects, turn out to be significantly weaker than the U-test results. See Table A.12 in the Appendix. For the vignettes, except for (e), the inclusion of additional variables renders the observed effects insignificant.

1.4 Summary

The seminal study by K. Chen (2013) shows that a person's future-oriented savings rates is lower, and future-oriented health behavior is less prevalent if a person's native tongue is classified as a strong-FTR language, i. e., it demands the use of a grammatical future tense when describing future events. As word meanings and connotations are not perfectly aligned across languages (Thompson et al., 2020), studying FTR across languages can introduce confounds, making a clean identification of an FTR effect spurious. To avoid such potential confounds, we implemented a within-language experiment with German native speakers. The German language allows for referencing future events in the grammatical present as well as the future tense, facilitating a within-language investigation. We tested the impact of present-tense and future-tense framing in the German language on behavior related to time and risk preferences in a medium-scale online framing study. We included two sets of vignettes to investigate the influence of

the present and future tense on subjects' beliefs about the immediacy and likelihood of future events. We further designed an exercise to elicit subjects' preferences for a specific tense to reference future events and supplemented the study with a socioeconomic survey. Our within-language variation of future-time referencing (FTR) and experimental setting allows for a clean identification of the impact of only varying the grammatical tense in the German language on economic behavior and beliefs.

Our results show no significant effect on people's behavior in the risk and time preference elicitation tasks. However, we find evidence for influences of future and present tense reference to future events on subjects' beliefs. Nonetheless, the effects are weak and can be counteracted when additional preconceptions are introduced.

Our experiment finds no evidence in support of the idea that altering the grammatical future tense reference impacts economic decision-making in a meaningful way. Consequently, when viewed in conjunction with the results of K. Chen (2013), our results hint at a deeper but imperfect link between language and culture (and maybe even subcultures). That difference is not necessarily captured in a simple culture dummy. We therefore argue that future, cross-border research should take this into account by relying on additional language and language-use information, as well as between language alignment, in order to more cleanly control for culture effects when testing behavioural channels.

CHAPTER 2

Charting the Type Space

The Case of Linear Public-Good Experiments

Abstract: Behavior in economic games is not only noisy. One has reason to believe that heterogeneity is patterned. A prominent application is the linear public good. It is widely accepted that choices result from participants holding discernible types. Proposed types, like free-riders or conditional cooperators, are intuitive. But the composition of the type space is neither theoretically nor empirically settled. In this paper, we leverage machine learning methods to chart the type space. We use a simulation to understand what can be achieved with machine learning. We rely on these insights to find clusters in a large (N = 12,414) set of experimental data points from public-good games. We discuss ways in which these clusters could be rationalized. Finally, we offer two outlooks, one traditional economic approach and another rooted in supervised learning, on how to go forward with our results.

JEL Codes: C71, H41

2.1 Introduction

Standard theory predicts the tragedy of the commons. Everybody maximizes individual profit, and exploits socially minded choices of others. If members of the community interact repeatedly, but it is known when interaction will stop, the gloomy prediction still holds. A robust experimental literature shows that, in the aggregate, results look different. In a standard symmetric linear public good, average contributions typically start considerably above zero, but tend to decline over time (J. O. Ledyard, 1995; Zelmer, 2003; Chaudhuri, 2011). A substantial theoretical literature rationalizes these results, usually by introducing some form of social preference into the utility function (for an excellent overview, see Fehr and K. Schmidt, 2002). While such extensions of motives can generate a starting point above zero, it is more difficult for them to explain the downward trend also. For this, one needs a reactive element. It has been prominently introduced into the literature with the concept of conditional cooperation (Fischbacher, Gächter, and Fehr, 2001). A conditional cooperator is willing to act unselfishly provided she expects or knows that others will do so as well. In principle, the downward trend could result from the fact that conditional cooperation is imperfect. While participants would not be outright selfish, they would still try to outperform their peers, albeit only slightly (Fischbacher, Gächter, Bardsley, et al., 2010). Engel and Rockenbach (2020) have shown that this explanation is not supported by the data. Rather, the downward trend results from bad experiences. If participants, in the previous period, have been overly optimistic about the contributions of their peers, they adjust their beliefs and, in turn, their contributions in the subsequent period. Critically, they overreact to negative experiences.

This is where the present project starts. If the population were homogeneous and completely consisted of conditional cooperators, there could not be a downward trend. The source of the trend, and hence the need for at least some form of institutional intervention to sustain cooperation, must be heterogeneity. Even if many individuals are in principle good-natured and happy to cooperate in good times, their willingness to do so is fragile. If they experience exploitation, they react. While the claim is intuitive that populations are heterogeneous, understanding the character of this heterogeneity is inherently difficult. One needs estimates about the utility functions of group members: is an individual outright selfish? Is she so strongly motivated by the common good that she does not care about the choices of others? Or does she react? If so, what does she react to? And how strongly? There could also be mixed types: individuals free-ride or cooperate for that matter, unconditionally, as long as a certain threshold is not crossed. Reaction functions might have an exploratory component: while an individual is in principle of a certain type, she occasionally tests the waters by contributing more or less than suggested by her ordinary reaction function. Reaction functions could be non-linear. Conditional cooperators might, for instance, be happy to tolerate an occasional bad experience (maybe attributing it to others having made a mistake), but they might lose faith and react very strongly if bad experiences repeat. There might be individuals who try to educate their groups by showing them what could happen if others do not stop

misbehaving. For that purpose, they might once contribute nothing and go back to high contributions in the following period. Reaction functions may also depend on the effects of occasional exploitation. In the standard setting (group size 4, marginal per capita rate .4) three loyal members still make a small profit if they continue to cooperate (and accept that the free-rider gains a windfall profit).

All these behavioral programs resonate with data from public good experiments. But these are only ex-post rationalizations. Moreover, not every dataset could be reasonably explained with all of these behavioral programs. Before the field can move forward, and better targeted interventions can be designed, one needs a much deeper understanding of behavioral heterogeneity. Ultimately, it would be highly desirable to define formally, and experimentally test, these reaction functions. But a necessary first step is exploratory: which reaction functions exist, and how prevalent are they? Charting the type space is the aim of the present project. We start from the assumption that the theoretical possibilities for the composition of the type space are at best partly understood. We further note that reactions may not only differ in kind, but also in degree, which is why parameters must be estimated. This is why we revert to machine learning. We use a reasonably large dataset of earlier linear public-good games to find types, and discuss reaction functions that would rationalize the reaction patterns.

In principle, choice data are well suited for our endeavor. The choices of others in previous periods are the only information to which participants can react in an anonymous linear public good. For each individual, we can check whether and, if so, in which ways they have reacted to past choices of the remaining members of their group. We can represent the development of their choices over time as a time series. We can use the rich set of methods developed in the machine learning community for clustering the time series of choices, giving the algorithm the possibility to use the average choices of the remaining group members in the previous period as an input. From these clusters, we can extract what machine-learners call a prototype.

This approach, however, presupposes that reaction functions can indeed be inferred from choices. Arguably, this will depend on at least two features of the data: the precision with which an individual participant has reacted to experiences, and the character of these experiences. The former depends on the noise rate. Potentially, individuals have a certain reaction function, but they do not act upon it at all times. The latter depends on group composition, and on initial choices. To illustrate: in a group of three straightforward free-riders, a conditional cooperator can be expected quickly to make choices that are indistinguishable from the choices of native free-riders. Discriminating between the choices of conditional cooperators and of free-riders will be the more difficult the lower the initial contribution of a conditional cooperator is. It should be equally difficult to discriminate between conditional cooperators and genuinely cooperative participants if a single conditional cooperator is surrounded by a group of native cooperators.

Before using machine learning for clustering participants in real data, we therefore investigate with simulated data the framework conditions under which potentially powerful algorithms can find types. In simulations, we can systematically vary the composition of the type space, the definition of individual types, and the noise rate. This first step yields

one important insight: machine learning methods find patterns. If the choice program of an individual is reactive, one and the same choice pattern may result from different reaction functions, depending on the choices the remaining group members have made in the previous period. Consequently, there is no one-to-one mapping between patterns and types. This must be reflected in the design of the clustering algorithm. We show that interpretation becomes much easier if one estimates a number of patterns that is considerably bigger than the expected number of types and hence reaction functions.

Simulation also helps us with two further tasks. We can estimate the richness of the data that is required for making the exercise meaningful. And we can check in which ways fine-tuning the algorithm improves estimation.

As explained above, we do not take it for granted that the type space has already been understood completely. A major motive for our project is the possibility that there are further types that have not been theorized. Yet, for our simulations, we need to build in types that have already been conceptualized. In the simulations, we work with groups consisting of different fractions of the following five types: The first are altruists, whom we define as participants who do not react to experiences, and who start with relatively high contributions. Such participants may exhibit variance, and all the more so the higher the noise rate. But they show no trend.
The corresponding type at the lower end is total free-riders. They in principle do not make contributions to the public project, but may occasionally deviate from this program.
Pure conditional cooperators start with relatively high contributions, but adjust them to experiences.
Following Fischbacher, Gächter, and Fehr (2001), we allow for hump-shaped contributions: up to a value near half the endowment they increase contributions in reaction to good experiences, but they exhibit a perverse reaction to even better experiences.
Following Engel and Rockenbach (2020), we finally implement farsighted free-riders. For some initial periods, “they feed the cow” by making substantial contributions, but then start “milking” it by reducing their contributions below average contributions in the previous period.

The remainder of this paper is organized as follows: In Section 2.2, we situate our endeavor in the literature. A small number of types have already been theorized. We use these types to simulate data in section 2.3. We use this dataset for two purposes: In Section 2.4, we show why a naive approach cannot work: once one allows for choice functions (of at least some types) to be reactive, there is no one-to-one mapping between types and what one can observe in the data, i.e., choice patterns and their corresponding experience patterns. In Section 2.5, we use theory and an extended grid search to find the best algorithmic configuration for clustering this kind of data. This prepares the main Section 2.6, where we apply the method to a sizeable set of experimental data. It turns out that empirical choice/experience patterns are much richer and quite different from the patterns resulting if one exclusively assumes types that have already been theorized. Section 2.8 concludes with discussion.

2.2 Literature

It has often been noted that choices in public good experiments are not homogeneous (see only Fischbacher, Gächter, and Fehr, 2001; Fischbacher, Gächter, Bardsley, et al., 2010). But the literature has only relatively recently begun to define the type space more precisely. Amin et al. (2018) use theory derived from Fischbacher, Gächter, and Fehr (2001) to classify 72 participants from a new experiment into 7 types, and then use simulation to find out which fraction of which type is required to sustain cooperation in a linear public good. Lucas et al. (2012) show with simulation that cooperation is hard to sustain in a linear public good if the group consists of heterogeneous types (which they take from Fischbacher, Gächter, Bardsley, et al. (2010)). Arifovic and J. Ledyard (2012) develop a model that combines social preferences with learning. In the framework of this model, conditional cooperation is not a type, but develops endogenously. They use data from, among others, Isaac and Walker (1988) and J. Andreoni (1995) to calibrate their model, and argue that it has a good fit. We have a different goal. On the one hand, we do not expect individual choices to be merely noisy. We consider the possibility that heterogeneity is patterned. On other hand, we do not assume that the behavioral forces that drive this heterogeneity are already fully understood. Rather, we wish to find patterns that are hard to reconcile with extant theoretical concepts. The purpose of our exercise is hypothesis generation. Testing these hypotheses would require a series of new experiments. That is beyond the scope of the present paper.

Engel (2020) also uses machine learning to organize the type space for experimental data, demonstrating the approach with data from Fischbacher, Gächter, Bardsley, et al. (2010). Yet, he has a different research question. He wants to compare the performance of a finite mixture model (which estimates the type space and choices conditional on type simultaneously) with a two-step approach. First, he estimates the type space from the data, and then estimates choices conditional on type in a mixed-effects model. That model interacts the types estimated in the first step with the effect of experimental manipulations. He also uses a different approach for estimating types, using the coefficients of local (per participant) regressions as inputs for a classification and regression tree.

A third group of contributions is more remote. Game theory usually starts with a complete definition of the game which includes the strategies available. Yet, when they are exposed to one of the games of life, individuals often do not know that much. They must learn what game they are playing and what strategies are available. This task is even harder if they cannot exclude that the population they play with with is heterogeneous. However, games can be too complex for solving them analytically. Then solutions must be found computationally. Ficici et al. (2012) make the game tractable by first compressing a large number of agents into a manageable number of clusters, and then solve the simplified game analytically.

Closest in spirit are Bapna et al. (2004) and Y. Lu et al. (2016). Both papers aim at classifying bidding strategies in online auctions (Bapna et al., 2004) and in flower auctions (Y. Lu et al., 2016), using machine learning methods. Vorobeychik et al. (2007) use machine learning methods to find the strategy space of infinite games . Mao et al.

(2017) use experimental data from a prisoner's dilemma to specify a classic learning model that helps them to divide players of a prisoners' dilemma game into two distinct behavioral types. The main difference to us is that we look at a different, more complex game (a dilemma), to which prior results are not easily transferred. Moreover, we use experimental data and exploit the power of algorithms for the classification of time-series data.

2.3 Data-Generating Process

Linear Public-Good Games. While we believe our method to be applicable more generally for finding patterned heterogeneity in repeated, interactive experiments, our specific object of investigation is a linear public good. The game is defined by the following profit function,

$$\pi_{it} = e - c_{it} + \mu \sum_{k=1}^{K} c_{kt} \tag{2.1}$$

,where π is profit of individual i in period t. Every period, the individual receives an endowment e. She can keep the endowment or make a contribution c to the public project of the group. Marginal per-capita rate $0 < \mu < 1$ creates the dilemma. As $\mu < 1$, each individual is best off keeping the entire endowment for herself. Yet, as $K\mu > 1$, the group is best off if all members contribute their complete endowments. Most frequently, $e = 20, \mu = .4, G = 4$ have been chosen (J. O. Ledyard, 1995; Zelmer, 2003; Chaudhuri, 2011). Then, three loyal group members still make a small profit. This serves as a buffer against the rapid decline of contributions.

Simulated Type Space In their seminal paper, Fischbacher, Gächter, and Fehr (2001) argue that (in their one-shot version of this game) there are three types: free-riders, conditional cooperators, and "hump-shaped" players. In his reanalysis of Fischbacher, Gächter, Bardsley, et al. (2010), Engel (2020) further finds a small, but discernible fraction of altruists. In their reanalysis of Fischbacher, Gächter, Bardsley, et al. (2010), Engel and Rockenbach (2020) use a combination of belief and choice data to distinguish a fifth group, which they call far-sighted free-riders. In our simulations, we allow for these five types. We focus on a partner design. Groups stay together for the full duration of the game. We always allow for an individual random effect η_i and residual error $\sigma_{it} \perp \eta_i$, which we both define to be normally distributed with mean 0 and standard deviation .3 ($\sim \mathcal{N}(0, .3)$). We thus implement the type space as defined in Table 2.1, where $c_{-i,t-1}$ is the average contribution of the remaining group members in the previous period $p-1$.

We have groups of size $G = 4$, and we allow for $t = 5$ types. Participants choose their contributions to the public good simultaneously, which is why their order does not matter. We consider the possibility that types are present more than once in a group. Hence, we have a problem of unordered sampling with replacement. This gives us a total type space of

Table 2.1: Simulated Type Space

type	$p = 1$	$p > 1$
Short-sighted free-rider	0	0
Far-sighted free-rider	10	$c_{-i,t-1}$ if $t < \tau$ 0 if $t \geq \tau$
Conditional cooperator	10	$c_{-i,t-1}$
Hump-shaped	5	$c_{-i,t-1}$ if $c_{-i,t-1} \leq 10$ $-c_{-i,t-1}$ if $c_{-i,t-1} > 10$
Altruist	20	20

$$N = \binom{t+G-1}{G} = \frac{(5+4-1)!}{(5-1)!4!} = 70 \tag{2.2}$$

different group combinations. In our simulations, we include each of these 70 combinations of types 4 times. As three of the five types (conditional cooperators, far-sighted free-riders, hump-shaped players) are reactive, we give the classification algorithm access to the exact same experiences that participants make in this design, i.e, the mean contribution of the remaining group members in the previous period. Hence the object of clustering is a two-dimensional time series consisting of the own contributions over time, as well as the lagged past experiences in terms of average contributions within the group over time. We run the simulations for different number of periods $p \in 10, 15, 20, 25, 30$. The results do not differ with the number of periods P.

2.4 The Naive Approach

Confusion Matrix. Simulation is routinely employed to test the performance of an estimator. One generates a dataset where one knows ground truth and checks whether a proposed estimator reconstructs the simulated parameters reasonably well. If an alternative estimator outperforms a competing estimator, one adopts the better-performing method. Simulation gives the researcher confidence in the use of an estimator with data where she does not know ground truth.

When applied to our estimation problem, the seemingly straightforward criterion for choosing an estimator would be the frequency of identifying the simulated types. Assessed with this criterion, the results reported in Table 2.2 are sobering.[1] Each of the 5 types is present exactly 224 times in the dataset. Yet, the size of the clusters ranges from 92 to 400. All clusters except the third are fairly impure: participants from different simulated types are put into the same cluster. Even knowing ground truth, it is hard to match clusters with types. Cells are highlighted in green if at least the most frequent type per cluster and the most frequent cluster per type coincide. In the example dataset,

[1]For consistency, we use the same algorithmic configuration that we develop in Section 2.5 and that later apply to the experimental data in Section 2.6.

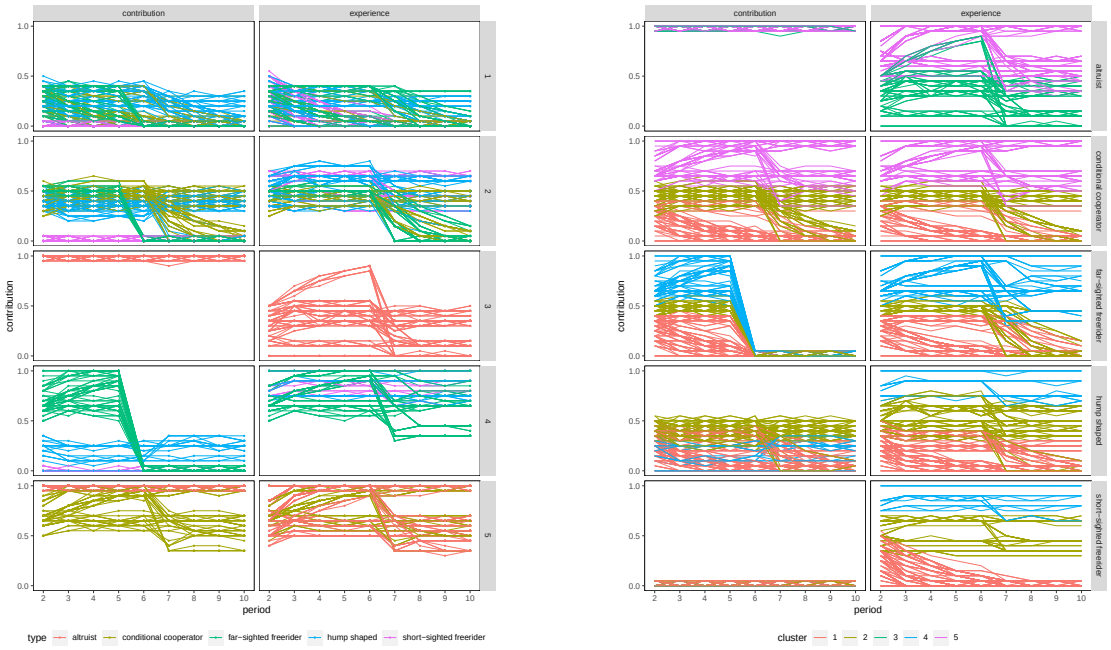
this only holds for the two non-reactive types: altruists and short-sighted free-riders. But even the purity of these two clusters is low. In cluster 5, 33% are actually conditional cooperators. In cluster 1, only 35% are indeed short-sighted free-riders. The remaining 65% consist of 27% hump-shaped types, 22% far-sighted (and hence partly reactive) free-riders, and 17% conditional cooperators. For all reactive types, one needs secondary (hump-shaped types, yellow cell) or tertiary (far-sighted free-riders: red cell) criteria for matching clusters with types. For cluster 3, no unique type can be found (as altruists are even more prominent in cluster 5). This is why one cannot even match the highest frequency in the cluster with the highest frequency in a type if one no longer considers clusters and types that have already been matched in an earlier round of matching.

Table 2.2: Confusion Matrix

cluster	1	2	3	4	5	Total
Altruist			92		132	224
Conditional cooperator	68	92			64	224
Far-sighted free-rider	86	74		64		224
Hump-shaped	106	94		24		224
Short-sighted free-rider	140	68		16		224
Total	400	328	92	104	196	1120

Clusters Are Patterns, Not Types. Figure 2.1 shows why the attempt fails to validate 5 clusters by comparing them with 5 simulated types. The algorithm does a reasonably good job at clustering the data. But it clusters patterns of observed contributions, combined with patterns of observed experiences in past rounds (henceforth experiences). There is no one-to-one mapping of 5 patterns to 5 types.

Cluster 3 is the only pure cluster. It is defined by contributions being high, irrespectively of experiences. These altruists do even accept outright exploitation. The remaining altruists are in cluster 5. By definition, their own contributions are also at the top. But now experiences are more favorable. This is why the algorithm lumps altruists together with conditional cooperators. As they are reactive, in a substantial number of instances within this cluster, contributions drop in the middle of the time series. The kink, of course, results from the presence of far-sighted free-riders who start cashing in. Many of the far-sighted free-riders are put into cluster 4. They are together with hump-shaped players and short-sighted free riders, who both make low contributions throughout the game. Apparently, the decisive feature for putting a participant into this cluster is not her own contributions, but the contrast between low contributions (at least for some part of the time series) and considerably more favorable experiences. By the same token, clusters 1 and 2 are distinguished. In both clusters, contributions are rather low. However, in cluster 1, experiences are low as well, while they are discernibly higher than contributions in cluster 2. It is even more instructive to consider which types are put into which clusters, (Figure 2.1b) in the Appendix. Altruists, conditional



(a) 5 Clusters for 5 Types

(b) 5 Types in 5 Clusters

Figure 2.1: Clusters versus Types

cooperators and far-sighted free-riders are split into clearly distinct subgroups. In the case of altruists, the critical feature is experiences. If experiences are good, they are in cluster 5. Otherwise, they are in cluster 3. For conditional cooperators, the match between the level of their own contributions and experiences is decisive. If both are high, they end up in cluster 5. If both are low, they end up in cluster 1. If both are in the intermediate range, they are put into cluster 2. The same logic applies to far-sighted free-riders. They are in cluster 2 with intermediate and in cluster 1 with low contributions and experiences. Yet, if both contributions and experiences start high, they are not put in cluster 5, but in cluster 4. In principle, this is also the logic for hump-shaped players. If contributions and experiences are low, they are assigned to cluster 1. If contributions and experiences are intermediate, they are assigned to cluster 2. The only difference results from perverse reactions if experiences are too good, so that participants react by reducing their own contributions. These participants are put into cluster 4. Finally, by design short-sighted players cannot be distinguished by their own contributions. In the same way as hump-shaped players, they are distributed across clusters 4, 2, and 1, depending on the level of contributions by the remaining group members.
Hence, upon closer scrutiny, there is no problem with the performance of the algorithm. It just does not do what one might have naively expected. The object of classification is not types, but time series. Three of the types that we have simulated are reactive themselves. Unless the environment exclusively consists of short-sighted free-riders or

altruists (which only holds for 2 of 70 simulated group compositions), individuals with a consistent reaction function respond to a variety of environments. If we impose 5 clusters, the algorithm must distribute pairs of experiences and choices across these clusters as best it can.

If one allows for types to be reactive, one cannot directly infer reaction functions from the data. Precisely because types are allowed to be reactive, one and the same reaction function may lead to distinctly different choice patterns. Actually just considering choice patterns would be misleading as well. One would miss the possibility that, in certain environments, multiple types exhibit very similar behavior. In Figure 2.1a, the point is most forcefully illustrated by the biggest cluster, cluster 1. Since overall cooperativeness is low in these groups, the choices of conditional cooperators, hump-shaped players, far-sighted free-riders, and short-sighted free-riders look very similar.

One needs an indirect strategy if one wants to infer potentially reactive types from the data. The proximate object of discovery cannot be types. It must be two-dimensional patterns, i.e, combinations of the development of experiences over time with the development of choices over time. The data can only inform the researcher about the distinct characteristics of these patterns. As the next step in the research process, she must attempt to rationalize these patterns. That point is best illustrated by Figure 2.2. Here, we see the result of expanding the number of clusters. Immediately, the algorithm is able to find "pure" clusters. That becomes even more prevalent when comparing it to the clustering result in Figure 2.1.

In Section 2.5, we discuss alternative approaches for this task and define our preferred algorithmic configuration. In Section 2.6, we apply this approach to the experimental data.

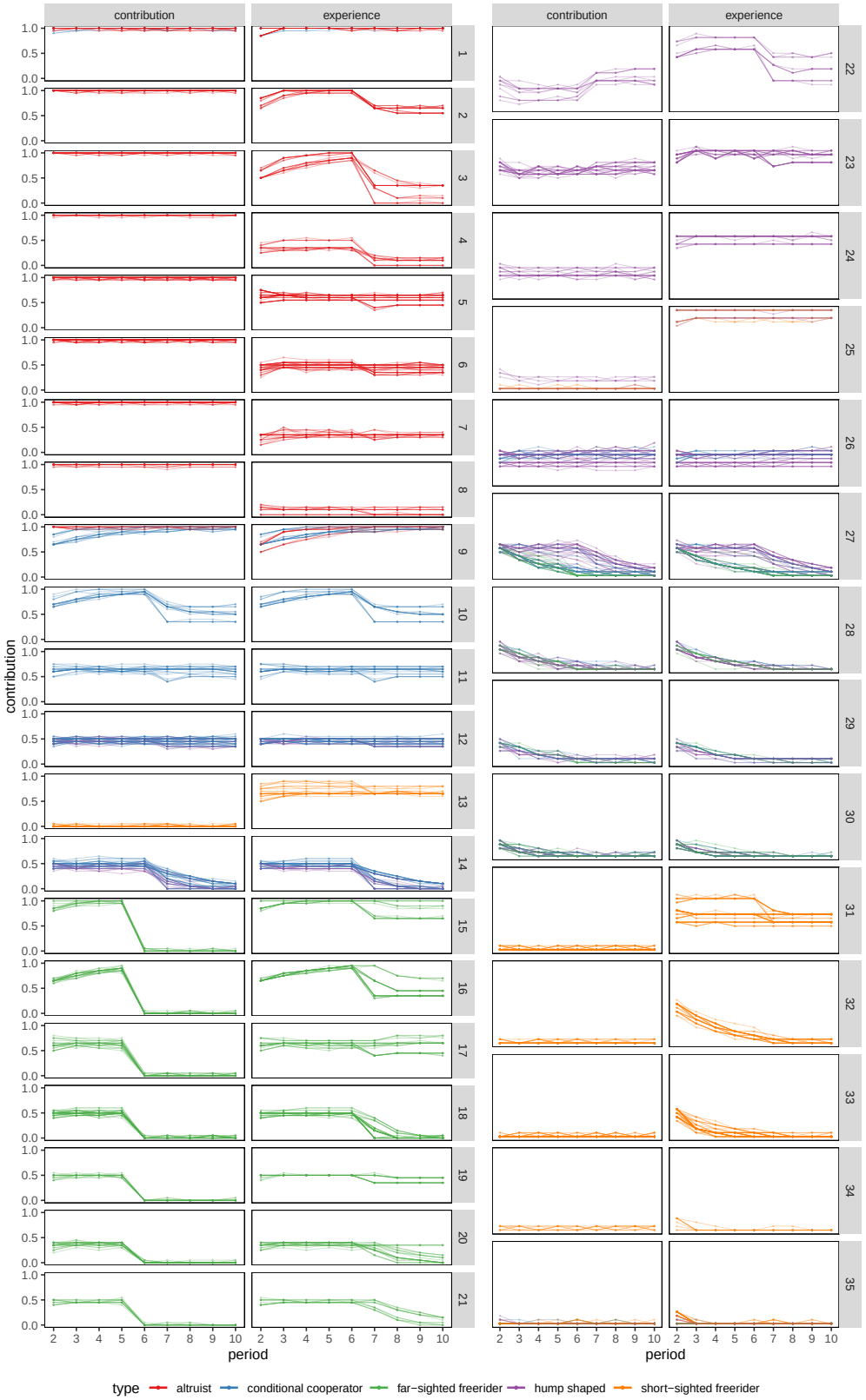


Figure 2.2: Exemplary Partitioning of the Simulated Dataset into 35 Clusters for 5 Types

2.5 Method

Clustering time-series data. Repeated experiments produce time-series data. It is meaningful to relate the choices of an individual at a given point in time to the choices this individual has made at an earlier point in time, and which she will make at a later point in time. From the development of choices over time, one can infer the program this individual has followed. In principle, one could capture the dependence of choices over time with the help of parameters of an appropriate transformation, and then cluster individuals with classic algorithms for static data (Liao, 2005); this is how Engel (2020) proceeds, using the coefficients of linear local (per participant) regressions as input for the classifier. This straightforward approach may well be sufficient for many practical applications. Yet, the approach requires that the local regressions adequately capture the characteristics of an individual's choice program. As in this project we want to find the best way to characterize these programs, we prefer a classifier that remains open to unexpected features of the individual time series. This is why we work with the raw time series, and we use algorithms that have been specifically developed for time-series data (for overviews, see Liao, 2005; Sardá-Espinosa, 2017).

Multivariate Clustering. Actually, many standard experiments are not only repeated. They are also interactive and produce panel data. In an interactive experiment, the program of an individual participant may react to the experiences she has made with the choices of others. This may hold for a cognitive reason: the individual learns from others; or for a motivational reason: the individual wants to react to the choices of others. In principle, the reactive component of the individual choice program could be captured by regressing individual choices on the experiences resulting from the choices made by other group members. Yet, this approach assumes that the reaction to experiences stays consistent over time. We are, obviously, open to this possibility, but do not want to impose it by the design of our estimation. This is why, instead, we provide the algorithm with the exact information that participants receive in the experiment. It consists of the average choice of the remaining group members in the previous period. The algorithm thus simultaneously receives two time series: the development of the choices over time that each participant has made; and the corresponding development of the average choices made by the remaining group members in the respective previous period.

Choice of the Clustering Algorithm Several methods have been developed for clustering (raw) multivariate time series, and they all come with multiple degrees of freedom (for overviews, again see Liao, 2005; Sardá-Espinosa, 2017). For our purposes, we need a clustering algorithm that is able to deal with multivariate data. In principle, this algorithm could be either hierarchical or partitional. In general, hierarchical approaches are preferable if one has reason to believe that the type space exhibits a discernible structure. This is not the case with our data, which is why we use a partitional algorithm (Hastie et al., 2009, chapter 13).

Cluster Evaluation. The number of clusters k is a free parameter. In order to select the best k, one has to use cluster validation indices. As our clustering problem is unsupervised, we have to rely on internal cluster validation indices. The following validation indices are well-established in the literature:

- Silhouette index (`Sil`)[2]
- Dunn index (`D`)
- COP index (`COP`)
- Davies-Bouldin index (`DB`)
- Modified Davies-Bouldin index (`DBstar`)
- Calinski-Harabasz index (`CH`)
- Score Function (`SF`)

These CVIs differ in the emphasis they put on cluster cohesion over cluster separation; whether they combine parameters by way of summation or division; whether or not they rely on normalization (for details, see Arbelaitz et al., 2013). As we have no strong conceptual reasons to prefer one CVI over the other, we employ all methods and aggregate over the outcomes.[3]

For simplicity, in the literature one picks either on or two CVIS without any specific criteria as to why, or else the choice is often made by majority vote. The former would seem arbitrary, yet, for several of the choices that we have to make, the majority vote is inconclusive. We therefore proceed the following way: for each choice parameter in question, we rank the scores of each CVI. For each outcome, we calculate the sum over all 7 ranks. We choose the parameter that receives the highest sum of ranks.

Selection of the Optimal Range for K. Section 2.4 makes it clear that we have to expect more patterns than types, and hence should estimate a number of clusters that is larger than 5. But which is the optimal number? As we know the data-generating process, for the simulated data we can derive the maximum from theory. In the dataset, we have 5 types who interact in groups of 4. From (2.2) we know that this leads to 70 distinct group compositions. One might think that the number of environments that a player may face is smaller, as there are only 3 others in the group. Yet, others are themselves potentially reactive. Then, the choices the individual in question has made in the past have shaped the experiences others have made in previous periods, to which they have reacted in turn. Hence, theoretically, there are 5 types $\cdot$ 70 environments = 350 different patterns. Imposing so many clusters would almost surely lead to overfitting. To strike a

[2]Letters refer to the code in `R` package `dtwclust`.

[3]As the clustering algorithm has a random starting point, we repeat the comparison with 15 different starting points and use the mean index per CVI. Three of these indices (COP, DB, and DBstar) are to be minimized. For comparability, we invert the scores of these CVIs.

balance between overfitting and underfitting, we proceed in two steps: In the first step, for a given dataset, we only consider any k for which the within-cluster variation ssw is $25\% \leq ssw \leq 10\%$ of the respective maximum and minimum. In Figure 2.3, we apply this method to the simulated data. As one sees, the range for k is within a sensible margin: considerably greater than 5, but much smaller than the upper bound of 350.

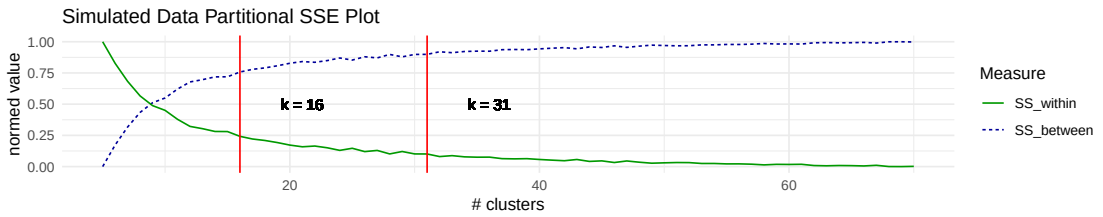


Figure 2.3: Simulated Data: Acceptable Range for k

Within the range thus defined, we then select the optimal k, using the cvi-ranking method introduced above.

Distance Measure. We further have to define the distance measure. For the clustering of time series, traditional measures like Euclidean distance are inappropriate, as they would overly depend on local differences. The most popular alternative is Dynamic Time Warping (DTW). It can capture similarity even if one time series is slightly shifted or has a slightly different shape (Berndt and Clifford, 1994). Yet, DTW is computationally costly. The procedure may occasionally even lead to pathological matches. Both concerns motivate the imposition of constraints. They limit the area that can be reached by the algorithm. We consider the GAK and the "soft DTW" (sDTW) constraints (Cuturi, 2011; Cuturi and Blondel, 2017).

Centroid. We finally consider two methods for defining the centroid of the respective cluster. With Partition Around Medoids (PAM), the centroid always is an existing time series, while DTW Barycenter averaging (DBA) constructs a synthetic centroid, which makes the method more robust (Petitjean et al., 2011).

Parameter Search. For finding the best specifications in distance measure, smoothing parameter γ when distance is sDTW as well as centroid function, we generate a grid of 4,608 simulated datasets. We allow for individual specific error $\eta \in \{0.6, 0.7, 0.8, 0.9\}$, residual error $\sigma \in \{0.6, 0.7, 0.8, 0.9\}$, sample size $N \in \{1, 2, 3, 4, 5, 6\} \cdot 70$, i.e, the full set of possible type combinations, and the window within which dynamic time warping is executed $w \in \{1, 2, 3\}$. For each point in the grid, we run the clustering algorithm one time for each possible variation of parameters, i.e, 9 times.[4] The best-performing variation for each grid point is then selected according to the rank-sum method outlined in the previous section.

[4]For a detailed description of all variations, please see section B.1

We have the following results: sDTW is always the preferred distance measure. Which smoothing parameter γ is optimal depends on the number of clusters k, as well as the size of the dataset N. For larger numbers of clusters, i.e, $k \geq 35$ and a larger dataset, lower smoothing in the range $0.007 \leq \gamma \leq 0.085$ is preferred. For $k < 35$ and a smaller dataset, a smoothing of $\gamma = 0.01$ is preferred. For the majority of all cases, the preferred centroid function is DBA, irrespective of the remaining parameters. Consequently, we use these parameters for the partitional algorithm.

2.6 Experimental Data

Section 2.4 has demonstrated in which ways, in a linear public good, a pair of two time series is related to the reaction function of a participant. The development of choices over time must be seen in the light of the development of experiences this participant has made. As we have explained, there is no one-to-one mapping between this two-dimensional times series and the reaction function, and hence the participant's type. Yet, we have shown in which indirect ways the type can be inferred. As we expect the type space to be limited, we use clustering (of two-dimensional time-series data) to organize the evidence. This gives us a methodology for the ultimate purpose of writing this paper: we want to infer from clustering real, experimental data whether the true type space differs from, or is richer than, the five types that have already been established and theorized.

Table 2.3: Information on Experimental Studies Included

Study	Periods	Endowment	Group Size	MPCR	Subjects
Diederich et al. (2016)	7	40	10	0.3	360
Diederich et al. (2016)	7	40	40	0.3	200
Diederich et al. (2016)	7	40	100	0.3	500
Diederich et al. (2016)	7	1,000	10	0.3	50
Engel, Kube, et al. (2021)	10	20	4	0.4	96
Nikiforakis and Normann (2008)	10	20	4	0.5	24
Engel and Rockenbach (2020)	20	20	3	0.4	30
Kosfeld et al. (2009)	20	20	4	0.4	40
Kosfeld et al. (2009)	20	20	4	0.6	176

Data. Table 2.3 defines the dataset. We only use data from linear public-good games without any experimental intervention, i.e, data from voluntary contribution mechanisms. We have a total of 12,414 observations from 1,476 participants. Figure B.2 visually represents the dataset. On average, all experiments featured in the dataset exhibit the characteristic negative time trend. Yet, there is considerable variance. The level of cooperativeness is differently high. The decay in cooperation is differently steep. In one experiment, contributions are even almost stable over time. We see this variance as an

advantage. It gives us more scope for finding unknown reaction functions, in particular due to variance in the experiences participants have made.

Table 2.3 shows that the experimental studies exhibit unique characteristics. Simultaneously clustering the complete dataset would obscure these differences. The most critical parameters seem to be the number of rounds played, and the size of the group. Technically, the difference in the number of rounds could be normalized by way of linear interpolation. Yet, as we show in Figure B.3a and Figure B.3b, interpolation introduces artificial noise into time series that, otherwise, appear quite regular. To avoid such artefacts, we separately cluster the data for subsets defined by the length of the interaction and the size of the group.

Results. Figure 2.4, Figure 2.5, and Figure 2.6 display the resulting clusters by subset. Each cluster's prototype is highlighted in bold. The individual time series in the respective cluster are represented by thin grey lines. This also informs how many pairs of time series are in the respective cluster. Comparing Figure 2.4, Figure 2.5, and Figure 2.6, a first result is patent: the prototypes differ profoundly between the three subsets. The typespace is not only richer than extant behavioral theory; it is also conditional on the context defined by the respective experimental protocol. The most striking difference likely results from the size of the group. While groups had size 3 or 4 in the remaining experiments, in the experiments with $t = 7$, the groups had size $N = 10$, 40, or 100. Regression to the mean is the likely reason why experiences in all clusters with length 7 are nearly flat, and close to the middle of the range. By contrast, with $t = 10$ and $t = 20$, experiences exhibit much greater variance, both within and across clusters.

Short Panel, Large Group. We start interpretation with Figure 2.4, $t = 7$. For our purposes, the high degree of homogeneity in experiences is fortunate. We effectively see reaction patterns while holding constant that participants experience fairly homogeneous mean contributions of others. These experiences are mildly favorable: the level of contributions is in the middle of the range; they do not change much over time.

We only find few clusters with approximately unconditional choices. In cluster 1 and 2, choices are close to the top, while experiences are much lower. We can explain this pattern with unconditional altruism. In cluster 3 and 4, we find unconditional, short-sighted free-riders. Participants of this type always contribute 0, or (in cluster 4) close to it.

In contrast, strict conditional cooperators should precisely match the experiences they have made (provided they expect others to behave in the same way in the next period as in the present period). Cluster 5 can be rationalized this way. As experiences are so consistently close to the midpoint, there is no room for the behavior theorized as "hump-shaped". The only other previously theorized type that can be traced in the data is farsighted free-riders. This type invests in cooperation in early periods, and exploits others in later periods. This holds for clusters 6 to 10.

The remaining 26 clusters are hard or even impossible to rationalize with the behavioral programs hitherto discussed in the literature. Clusters 11-17 could be interpreted as

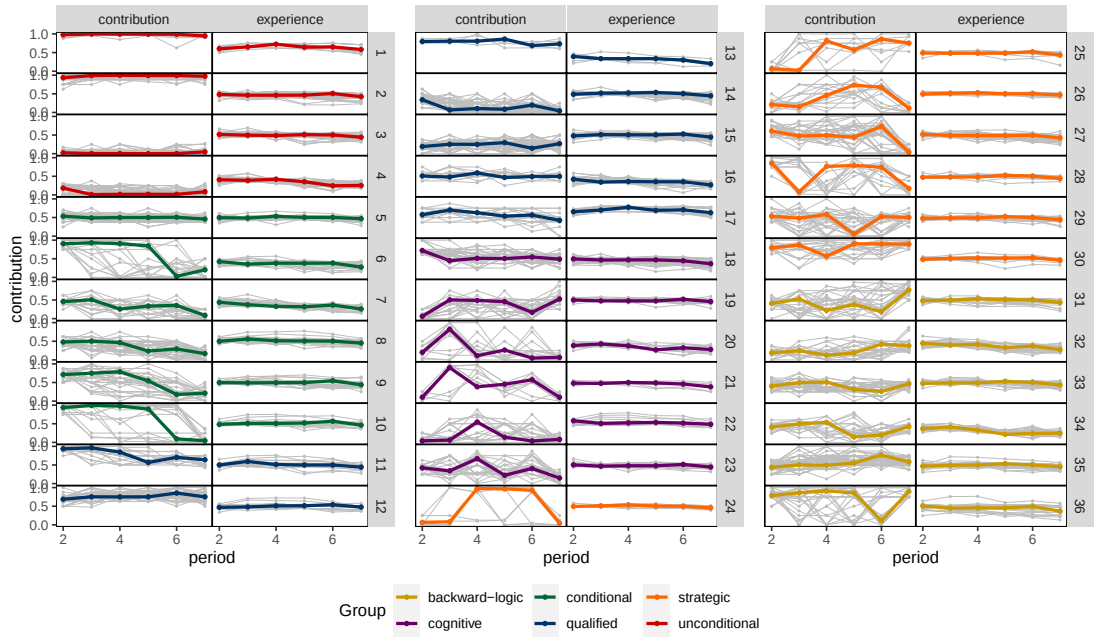


Figure 2.4: Cluster by Experimental Subsets, $t = 7$

qualified versions of known types. In clusters 11–13, contributions are substantially above experiences, but not at the top. This could be the behavior of an altruist who is, however, not willing to be completely blind to the choices of others. Clusters 14 and 15 are the mirror image at the low end. Contributions are not immediately and not completely at zero, but always below experiences. Finally, clusters 16 and 17 are imperfect cases of conditional cooperation. In cluster 16, contributions are consistently slightly above experiences, while they are consistently slightly below experiences in cluster 17.

Another potential behavioral program is of a cognitive nature. Participants are surprised by experiences and adjust their choices to the behavioral environment. This explanation is most intuitive in early periods, and provided the participant aligns her own choices with experiences. Clusters 18 and 19 closely fit this explanation. The participant had been either overly optimistic or overly pessimistic about the level of contributions. The remaining clusters with pronounced changes in the initial periods (clusters 19, 20, 21, 22, 23) require a more involved behavioral program. A consistent interpretation would be exploration. Exploration is reasonable if the participant in question not only has a reactive choice program herself, but considers the possibility that others have reactive programs as well. In that case,s she needs to test the waters and find out what is going to happen if she changes her own moves. The participant deliberately risks falling below the attainable period income as an investment into more profitable moves in the future. This explanation is particularly plausible for changes in early periods.

A participant who engages in (potentially) costly exploration can be said to act strategically. But in this interpretation the strategy is confined to making a better-

informed decision herself in future periods. The choice patterns in clusters 24–30 suggest a more encompassing strategic motive. The participant in question not only aims at optimizing her own future choices. She intends to induce other group members to behave in a way she considers more appropriate. In cluster 24, the participant seems to try leading by example. In clusters 25–27, the participant also, at least in some periods, contributes more than the group average. This could be motivated by the aim of signaling good intentions and the possibility of a brighter future to the group. In cluster 28, the participant might want to combine a warning what could happen if others don't follow suit with a positive signal later on. In clusters 29 and 30, only negative signals can be found.

In the final group of six clusters, participants increase contributions in the final or the penultimate period. We thus find an inverse endgame effect. As the game has a defined end, such choices cannot be motivated strategically. Participants must have deontological motives. If they had contributed less than average in earlier periods, a consistent interpretation is repent, leading to (at least partial) compensation. In cluster 36, the opposite interpretation as punishment invites itself. In the remaining clusters 33–35, contributions had been at or even slightly above the group average. At some point, contributions go down, but they go up again. This pattern would be consistent with an expression of discontent.

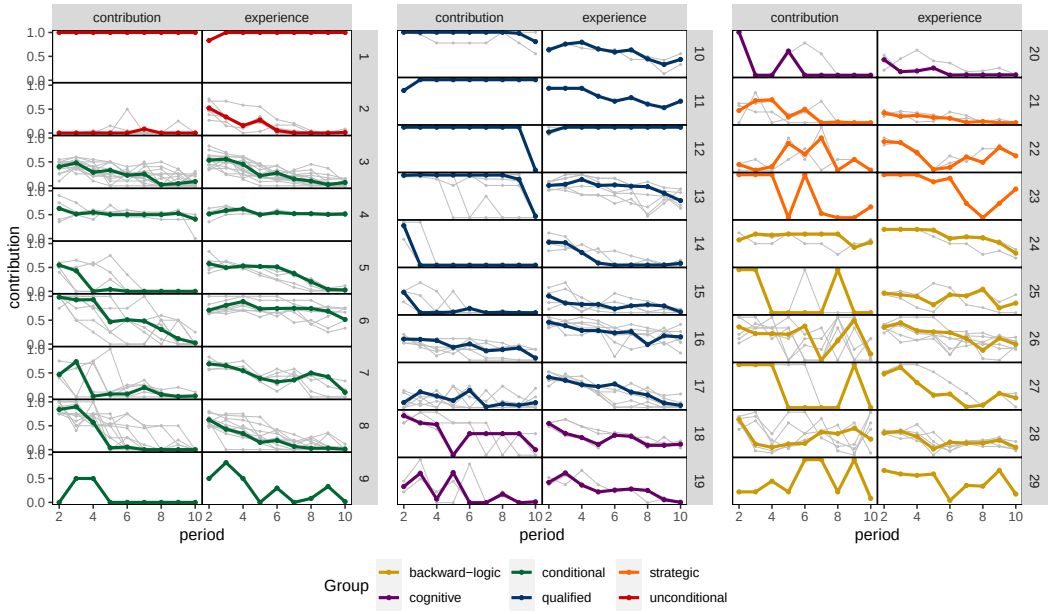


Figure 2.5: Cluster by Experimental Subsets, $t = 10$

Short Panel, Small Group. 10 periods are just three more than 7; yet, the patterns in the clusters for $t = 10$ – displayed in Figure 2.5 – look very different from the ones for

$t = 7$.[5] The obvious source of the difference is in the experiences participants are making. While in bigger groups regression to the mean conceals variance in the types of other group members, this variance may play itself out in the groups of 4, from which these data are taken. In many clusters we also observe the downward trend that has often been reported in the literature on public goods (see, e.g, clusters 3, 7, 8, and 17).

In this set of experiments, we find choice patterns consistent with unconditional altruism (cluster 1) and unconditional free-riding (cluster 2). But there are only very few observations in these two clusters. Again, not many clusters are such that contributions track experiences (clusters 3 and 4), which is what would be expected from a textbook conditional cooperator. A somewhat larger number of clusters is consistent with far-sighted free-riding, i.e, making reasonably high contributions in early periods, in the interest of cashing in at a later point (clusters 5, 6, 7, 8, and 9).

A further set of clusters are at best qualified versions of the previously theorized types. In clusters 10–13, in most periods contributions are at the top. But unconditional altruists would neither need a period to go to the top (had the participant initially been concerned about the degree of exploitation?), nor go down in the final period. Likewise, in clusters 14 and 15, contributions are low or even zero in most periods, but not in the beginning. Have these been far-sighted free-riders who do not consider investment in the corporation spirit worthwhile, given what they experience in the first period? Finally, in clusters 16 and 17, choices grosso modo track experiences, but at a lower level. Are these conditional cooperators intending to outperform the group at least slightly?

In clusters 18–20, we observe stark changes in early periods, either downwards (cluster 18), upwards (cluster 20), or both (cluster 19). These choice patterns could be motivated by exploration.

While exploration is also a possible interpretation in clusters 21–23, these patterns could also be motivated by the intention to educate the group, and hence to improve the outcome for all.

Finally in clusters 24–29, we see upward moves in the final or the penultimate periods, i.e, an inverse endgame effect. As explained with $t = 7$, one needs deontological motives to rationalize such choice patterns. In all clusters, at least for some periods, the participant had contributed less than the average in the previous period, and might feel moral urge to give back to the group, at least partially.

Long Panel, Small Group. The final set of observations is displayed by Figure 2.6 and comes from experiments with $t = 20$ and $N = 3$ or $N = 4$. Hence, in these experiments, participants stay together for a long time and therefore have many opportunities for learning from each other, and for attempting to influence each other. Both learning and influencing is meaningful, as there are only few interaction partners. The average contribution is a reasonable proxy for the composition of the type space. As Figure 2.6

[5] The total number of observations with $t = 10$ is much smaller than with $t = 7$. If we keep the upper limit of k at the level derived from the theoretical number of type combinations, we get too many clusters with very few observations. We therefore adjust the upper limit to $N/4 = 30$. With the help of the rank sum over all internal CVIs, we end up with the meaningful number of 29 clusters, depicted in Figure 2.5.

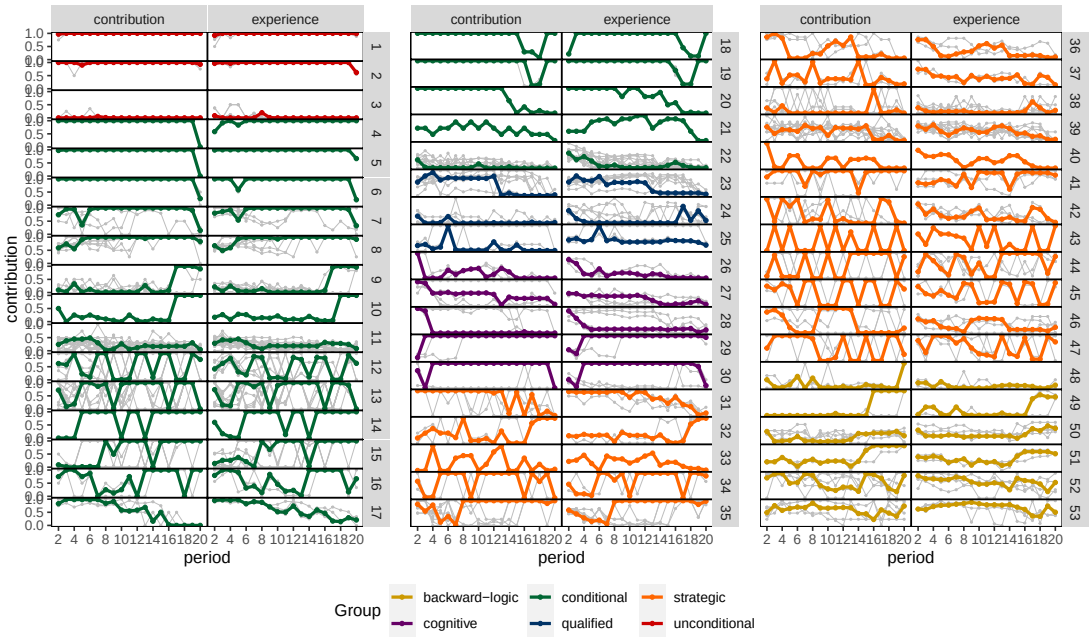


Figure 2.6: Cluster by Experimental Subsets, $t = 20$

shows, the different institutional setting again matters profoundly. In this context, experiences exhibit considerable variance. Choices quite often track experiences fairly closely.

Using the same set of cognitive and motivational effects as with $t = 7$ and $t = 10$, we can organize the type space. Yet, it is composed very differently. We do find a very small number of arguably unconditional altruists (clusters 1 and 2) and unconditional free riders (cluster 3). We also find clusters in which participants track the generally high (clusters 4–8) or low contributions of the remaining group members (clusters 9–11). The fact that there is at least some variance in experiences mirrored in variance in choices demonstrates that these participants are not qualified unconditional types, but react to what they observe.

The most characteristic difference between $t = 7$ and also $t = 10$, on the one hand, and $t = 20$, on the other, is the zig-zag pattern of both experiences and choices in many of the clusters. Apparently the longer time horizon has not helped, but hurt. Multiple opportunities for observing each other have made it difficult for many groups to get in sync. They have oscillated between low and high contributions, although there is no point in taking turns in a linear public good, which is a cooperation and not a coordination problem.

In clusters 12–17 choices and experiences match so closely that behavior fits classic conditional cooperation. By contrast, in clusters 18–22 it is the participant who triggers the download trend of the group, which suggests that these participants are far-sighted free-riders.

In clusters 23–25, choices slightly deviate from experiences, but no straightforward alternative interpretation invites itself. This is why we have classified these clusters as instances of qualified conditional cooperation.

With $t = 20$, we do not find many clusters where choices can be rationalized by a merely cognitive effect. In clusters 26–28, participants might have been too optimistic initially. In cluster 29, they might initially have been too pessimistic. Cluster 30 is the only cluster that suggests exploration in the first two periods.

In the bulk of not only clusters, but also data, participants seem to aim at inducing a higher contribution level in their groups. They either periodically go up, if not to the top. Or they signal that others should not take their benevolence for granted, by temporarily reducing contributions, frequently to 0 (clusters 31–47).

Finally, once again we observe five clusters with an inverse endgame effect (clusters 48–53). In cluster 50, contributions had clearly been below experiences for many periods. In this cluster, compensating the other group members for anti-social behavior is a consistent explanation. This explanation might also matter in clusters 52–53. By contrast, choices in clusters 48 and 49 look more like an expressive act, showing others how much more favorable outcomes might have been, had they been less selfish.

2.7 Rationalization

Section 2.6 has shown pronounced heterogeneity. It is beyond the scope of the present paper to theorize all the many patterns that we observe. The fact that we find discernible clusters gives us confidence that these patterns are meaningful. In the previous section, we have offered plausible interpretations. But these interpretations are, of course, only hypotheses. New experiments will be needed to isolate potential cognitive and motivational channels. In this section, by way of illustration, we zero in on one striking difference: with short panels and large groups (Figure 2.4) and with slightly longer panels and small groups (Figure 2.5), individual and group patterns are much smoother than with longer panels and small groups (Figure 2.6): in the data from games repeated for 20 announced periods, we find a lot more zig-zagging. Obviously, the longer shadow of the future has a strong behavioral effect. In this section we discuss an explanation for the observed sudden, drastic changes in behavior.

In a linear public good, payoff is given by (2.1). In the stage game, a participant maximizes profit by complete free-riding, i.e, by setting $c_i = 0$. As the linear public good is a prisoner's dilemma, $c_i > 0$ is dominated. If a selfish player assumes that all other group members are selfish as well, i.e, when assuming common knowledge of rationality, $c_{i,t} = 0, \forall t$ is the best response. This is the well-known unraveling prediction. As participant i expects all other participants j to choose $c_{j,t=T} = 0$ in the final period T, there is no scope for investing in cooperation in early rounds.

In their seminal paper, Kreps et al. (1982) have shown that this result breaks down when allowing for behavioral uncertainty. They show this for the belief that the counterpart (in a 2x2 prisoner's dilemma with discrete {cooperate, defect} action space) might either play tit for tat, or might be a conditional cooperator. Yet, in their model,

a longer shadow of the future is unequivocally beneficial. If T is large enough, in early rounds a rational player never defects. She would reveal that she is actually selfish. At best, her counterpart plays tit for tat, and punishes her for her deviation from the cooperative path of the group. If her counterpart is selfish as well, they end up in the the {defect, defect} equilibrium for all future rounds. We, by contrast, find the opposite. With a longer shadow of the future, there are more deviations. We also find upward deviations, not only downward ones. We thus need a different model to rationalize this observation.

In the initial step, we only allow for genuinely cooperative players. For this type $u_i(c_i < \bar{c}_j) < u_i(c_i = \bar{c}_j)$, where c_i is the contribution of player i to the public good, and $\bar{c}_j$ is the average contribution of the remaining group members to the public good. We work with this average as, in the experiments from which we have data, participants did not get feedback about individual contributions of the remaining members of their groups.

In principle, such a group is able to sustain cooperation. Yet, no player wants to be the sucker: $u_i(c_i = \bar{c}_j) > u_i(c_i > \bar{c}_j)$. When choosing how much to contribute in period t, participants do not know how much the remaining group members are going to contribute. They must work with the expectation $E(\bar{c}_{j,t})$. In later periods, they have a signal: $E(\bar{c}_{j,t}) \approx \bar{c}_{j,t-1}$. But in the initial period, they must work with their home-grown beliefs. Cooperation may fail. Not because other group members are genuinely selfish, but because at least some of them have been too skeptical initially.

Against this backdrop, it can be rationalized that a player sets $c_{i,t} > \bar{c}_{j,t-1}$: she uses this to signal her type. The signal is credible, as a player who is not willing to sustain cooperation has no reason to do that. The signal can be interpreted as an investment. The participant accepts to be temporarily exploited, in the interest of lifting the entire group to a higher contribution level. Such an investment is the more profitable the longer the shadow of the future is. This explains why such choices are more frequent in games with a larger number of periods.

The fact that $\bar{c}_{j,t=1} < e$, where e is the endowment, and hence the maximum contribution, may have more than one reason. Either all other group members were indeed conditionally cooperative, but too skeptical; or at least one of them was actually (shortsightedly) selfish. If the participant has invested in signaling her cooperative type in period $t = 2$, she must wait another period to learn. If $\bar{c}_{j,t=3} = e$, the group has coordinated at the maximum. In this logic, participant i not only sets $c_{i,t=2} = e$, but also $c_{i,t=3} = e$. She thus gives the other group members a chance to adapt to the cooperation signal she has sent in $t = 2$, and only reverts to $c_i < e$ in period $t = 4$ if $\bar{c}_{j,t=3}$ has proven that her strategy did not work out, as the group is actually not cooperative.

While this strategy is consistent, it puts a high burden on the group member who attempts to trigger the virtuous cycle. If she was too optimistic about the composition of the type space, she has to accept exploitation for two consecutive periods. Now her strategy is motivated by the possibility that an all-cooperative group is stuck in a bad equilibrium. The fact that she signals her cooperative type in $t = 2$ can be interpreted as the contribution to a second-order public good Yamagishi (1986) and Heckathorn (1989): one of the cooperative players must accept temporary exploitation for signaling her type.

This interpretation provides scope for an alternative strategy. The group member who has made the initial move, in $t = 2$, expects other group members to follow suit in $t = 3$. Hence, her strategy would be

- $c_{i,t=2} = e$
- $c_{i,t=3} = \bar{c}_{j,t=1}$
- $c_{i_t>3} = e|\bar{c}_{j,t=3} = e$; $\bar{c}_{j,t=1}$ otherwise.

Hence, this group member expects to be compensated in $t = 3$ by the remaining group members for her initial cooperative move, by them tolerating that she reduces her contribution to the original contribution level in the group, or even to 0. If cooperative participants use this strategy, we should see one of two patterns. In groups that are actually non-cooperative, we should see jumps to the maximum lasting one period which have no consequences at the group level. In groups that are actually cooperative, we should see such one-period-jumps to the maximum, followed by a jump downwards, followed by coordination at a high level. Hence we should observe zig-zagging.

The same strategy also works if the cooperative group member who takes the initiative is less optimistic about the composition of the type space. She may be open to the possibility that one or more of the remaining group members are actually selfish, but willing to sustain cooperation, as they expect the long-term profit from this strategy to be higher than early defection. Such players mimic genuinely cooperative players.

The player who takes the initiative may also be willing to tolerate partial defection. This is particularly plausible in the canonical design of the game, with 4 group members and MPCR = .4. Then 3 group members who cooperate fully still have a slightly higher payoff than from defection (24, rather than 20). Tolerating partial defection is also easier in the design of the game investigated in this paper, as participants only get feedback at the group level. They can therefore not see whether $\bar{c}_j < e$ is due to one player defecting, or all other players contributing less than the maximum, but the same amount.

Yet, if either possibility is taken into account, observing $\bar{c}_{j,t} > \bar{c}_{j,t-1}$ is less informative. Per se, genuinely cooperative players have no reason to revert to lower contributions in later rounds. By contrast a player who only mimics a cooperative type will start defecting once the expected payoff from defecting before others outweighs the gains from cooperation for future rounds. If other genuinely cooperative players are concerned about this possibility, they may themselves reduce contributions, as they lose faith in the willingness of others to cooperate. This concern looms even larger if cooperative gains are below maximum, as then gains from continuing to cooperate are smaller.

Taking these possibilities into account, we can also rationalize upward jumps in later rounds. By the same logic as in $t = 2$, a (genuinely or strategically) cooperative type wants to stabilize cooperation, by sending this cooperative signal. Again, the longer the time horizon, the more this strategy is profitable. And again, such a cooperative player may expect to be compensated in the subsequent period, by others tolerating her one-period defection. That is another way zig-zagging can be rationalized.

Once we allow for the belief that groups are heterogeneous, in the defined sense, we can also rationalize temporary downward jumps. This is straightforward if a player only mimics a cooperative type: she tests the waters. If others react, she knows that they are vigilant, so that (early) defection does not pay. Yet, a temporary downward jump can also be rational for a genuinely cooperative player who is skeptical about the motives of other cooperators. If they react by reducing their contributions, she knows that their cooperation is not genuine, and she can react by reverting to low contributions herself.

2.8 Discussion

The linear public good is one of the workhorses of behavioral economics. Hundreds of experiments have been run with this paradigm. The design is appealing as, in a stylized way, it captures what arguably is the essence of many conflicts of life, running from the degradation of the environment over the instability of a cartel to the precarious nature of any constraining institutional framework. The design implements a multi-person, multi-period prisoners' dilemma with a known end. If one assumes that actors exclusively maximize individual profit, the repeated game has a unique solution. In the final period, all group members will contribute nothing to the common project. Through unraveling, this is also the prediction for any earlier period.

The first experiments undertaken with this design have already refuted this prediction. On average, contributions start at some higher level, but decay over time. Per se, social preferences can rationalize positive contributions, but they do not predict the decay. Interestingly, per se the prominent concept of conditional cooperation cannot predict the decay either. If all group members are perfect conditional cooperators, and expect all others to follow the same behavioral program, any level of cooperation can be sustained, depending on initial beliefs. Fischbacher, Gächter, Bardsley, et al. (2010) propose a consistent explanation: the decay could result from conditional cooperation being imperfect. Participants would be willing to let themselves be guided by the level of cooperativeness in their group. But they would always try to undercut slightly. Yet, in their reanalysis of Fischbacher's and Gächter's data, Engel and Rockenbach (2020) have shown that true conditional cooperation is actually near-perfect. The decay results from heterogeneity. By the combination of choice data with belief data, they show that the decay results from the presence of short- and far-sighted free-riders. This is where the present project starts. It uses machine learning methods to cast light on this heterogeneity and chart the type space.

The paper makes a methodological and a substantive contribution. On the methodology side, it shows in which ways clustering can be used to infer the composition of the type space. On the substantive side, it shows that extant theories about behavioral types can only explain a very narrow fraction of the data.

Repeated experiments generate time-series data. In principle, the large family of algorithms for clustering time-series data are therefore appropriate. However, contributions could not exhibit a downward trend, unless at least some participants hold a choice program that is reactive. If we were to deprive the algorithm of the experiences participants

make, it would lump together choice patterns that are generated by completely different behavioral programs. This is why we use multivariate clustering and feed the algorithm with pairs of experiences and choices.

One might naively think that the algorithm will find as many clusters as there are distinct behavioral programs. With simulation, we show why this approach must fail. We simulate all combinations of five behavioral programs that have been theorized in the literature: altruists, conditional cooperators, far-sighted free-riders, hump-shaped contributors, and short-sighted free-riders. For investigating these five behavioral programs, we need many more clusters. Moreover, we also show that we do not need the theoretical maximum of 350 clusters; this would make the approach next to unusable for real data, as one would need a huge amount of data for that many clusters to be credible. We use internal cluster validation indices to find the appropriate trade-off between underusing and overusing the evidence.

We apply this methodology to a large dataset consisting of 12,414 observations. Results clearly show that the true type space is much richer than thus far assumed by the literature. Only a few of the clusters that we find in the experimental data can be rationalized with any of the five theoretical behavioral programs that we have used to simulate data. Obviously, the type space is considerably richer than typically assumed in the behavioral literature.

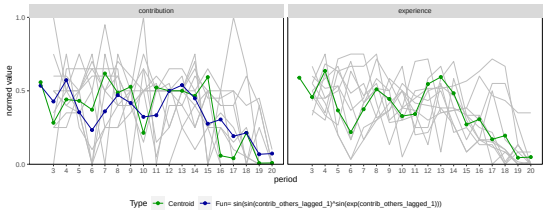
The main limitation of our approach is its exploratory nature. We alert the research community that the behavioral programs participants employ in an experiment, seemingly as simple as the linear public good, are very likely much richer, and much more heterogeneous, than typically assumed when designing and analyzing these experiments. This can obviously only be a first step. In the next step, frequent choice programs must be rationalized. In conclusion, we sketch two approaches that could productively prove complementary.

The first approach capitalizes on methods originating in physics. In physics (Udrescu and Tegmark, 2020), same as in industry (Francone et al., 1999; Castillo et al., 2002), problems are often too complex to start the analysis from first principles. Rather, one begins with the data and searches for succinct ways of rationalizing them. The approach is known as symbolic regression. Its main advantage is flexibility. Different from maximum likelihood, one need not impose functional form. Given (weaker) constraints on depth, a set of pre-defined functions from which the procedure may choose, as well as operators, symbolic regression aims to find the best combination of these inputs and the data, resulting in a functional form with the least error to the original values. Technically, symbolic regression is a supervised technique.[6]

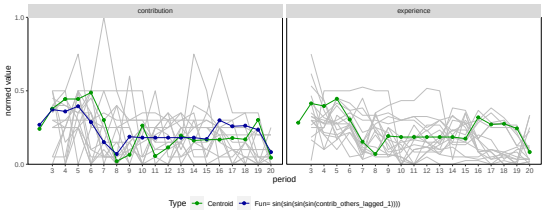
In the spirit of a proof of concept, in Figure 2.7 we show the resulting rationalizations for three selected clusters (from experiments with $t = 20$). As can be seen, the approximation works well. But the functional form that happens to fit the data not only differs widely across clusters; this could be necessary for defining the character of the heterogeneity. The functional form does not lend itself easily to interpretations either. If

[6]We make use of the R-package gramEvol (Noorian et al., 2016) which uses genetic programming (Kotanchek et al., 2003) to traverse the space of possible solutions efficiently.

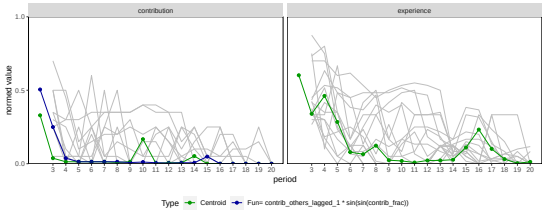
the collective endeavor of charting the type space progresses, one might be able to add meaningful explanations. But it might also turn out that symbolic regression is only good at prediction, not at explanation.



(a) Cluster 11



(b) Cluster 17



(c) Cluster 42

Figure 2.7: Exemplary Results Symbolic Regression

The alternative approach is more in the spirit of behavioral economics. In this perspective, one would read the evidence presented in this paper as a challenge for the development, and subsequent experimental testing, of more appropriate behavioral theory. One would aim at replacing the "asif" models from symbolic regression with process models. The interpretations proposed in section 2.6 offer building blocks for this enterprise. We conclude by explicating the ones that we deem most promising.

Unless a participant has an unconditional behavioral program, in the initial period she must work with beliefs about the behavioral programs that other group members are implementing. Participants might conceptually be allowed to condition their own program on the direction and the amount by which these beliefs turn out to be false. Additionally, participants might conceptually be allowed to invest in exploring the behavioral programs of the rest of their groups.

On the motivational side, one may enrich the concept of conditional cooperation. Participants would not blindly copy the experiences they are making. They would rather consider the possibility to influence the choices of other group members in future periods by their own choices in the present period. This could be theorized as costly signaling of type, including the intention to enforce a normative conviction in the future.

Besides such a forward-looking, strategic perspective, there might also be backward-looking motives. Participants might want to express their discontent and frustration, even if this reduces their own prospect for a higher profit. Or, conversely, they might want to reward others for unselfish acts. They might also, relatedly, repent their own

past behavior, and aim at partial reparation.

Sometimes, the next step forward in a line of research is not the final answer, but the right question. With the present project, we aim at demonstrating that heterogeneity in dynamic games, and in the linear public good in particular, is a promising frontier. This investigation is urgent if one hopes to learn from experimental data about the behavioral determinants of social dilemmas, in the interest of designing more powerful interventions.

CHAPTER 3

Text Classification of Ideological Direction in Judicial Opinions

Abstract: This chapter draws on machine learning methods for text classification to predict the ideological direction of decisions from the associated text. Using a 5% hand-coded sample of cases from U.S. Circuit Courts, we explore and evaluate a variety of machine classifiers to predict "conservative decision" or "liberal decision" in held-out data. Our best classifier is highly predictive ($F1 = .65$) and allows us to extrapolate ideological direction to the full sample. We then use these predictions to replicate and extend Landes and Posner's (2009) analysis of how the party of the nominating president influences circuit judge's votes.

Keywords: Judge Ideology, Circuit Courts, Text Data, NLP
JEL Codes: C8, K0

Supported by an IPAK-Travel Grant by Deutsche Forschungsgesellschaft

3.1 Introduction

In the United States, judges wield significant power due to the common law system (Dainow, 1966). The extent of U.S. judges' influence is a motivation for the extensive research into the determinants of judicial decision-making. In particular, there is a large literature on how opinions are affected by the ideology of the respective judge (e.g., Jeffrey A. Segal and Cover, 1989; Andrew D. Martin and Quinn, 2002; Andrew D. Martin, Quinn, and Epstein, 2004).

A leading paper in this literature is Landes and Posner (2009). This paper looks at how the party affiliation of U.S. Circuit Court judges affects the political ideology of their votes (conservative or liberal) on the court. While judges are nominally non-partisan, party affiliation can be proxied by the party of the appointing president or the party share in the Senate at the time of appointment. Landes and Posner (2009) show that judge party affiliation is statistically related to the ideological direction of votes. For their empirical analysis, they draw upon the Songer database of U.S. Circuit Courts,[1] which provides rich metadata, e.g., the political ideology of votes for each judge in each case. The classification of votes by ideological direction was a labor-intensive exercise which has led to frequent use in the empirical legal studies and political science literature (Ginn et al., 2015; Reid and Randazzo, 2016; Landes and Posner, 2009, e.g.).

Notwithstanding its broad use in the literature, the Songer database has some limitations. First, the political ideology classification has been assigned by human coders, which could be error-prone. These errors add noise to regressions and complicate replicability. In particular, as noted by Landes and Posner (2009), the political positions of conservative/liberal are not constant over time. Therefore, data coded in the past may not be categorized correctly, and Songer Project ideology labels for older Circuit Court opinions may be systematically incorrect.

Another problem with the database is the sampling approach. First, the database is only available for 1925-2002, so empirical analysis of vote ideology is only possible for that time period. Second, only a small set of cases was labeled (just 5 percent of the cases for those years). Finally, the authors used stratified sampling to get labels for similar numbers of opinions across courts and time. Therefore, the dataset is not representative of the full distribution of circuit court cases.

The goal of this paper is to address these shortcomings using machine learning and natural language processing techniques. The idea is to treat a machine to code the ideological direction of the votes. Within the set of labeled case, we can check how well the algorithm replicates human labels.

The classifier would provide a number of benefits. As soon as the classifier is trained, predictions even for an extremely large sample cost very little relative to hand-labeling (which require a human to read an opinion). We could potentially take the classifier to cases before 1925 and after 2002. Within the 1925-2002 period, we could classify the other 95 percent of unlabeled cases. Besides producing new labels, it could be used to audit and check existing labels for probable errors.

[1] The original, as well as the extended versions, are available at songerproject.org.

In this paper, we produce such a model. For the sake of interpretability, we focus on linear models. The model which worked best in our setting is a Ridge Classifier. Our model is trained on the complete opinion text in combination with the circuit, year as well as case type data. After optimization it achieves a cross-validated accuracy of 61.5% on the three label input and 66.5% on the two label subset. The final calibrated classifier working on the tow-label subset achieves the same accuracy score while increasing its precision as well its recall on the test set to 71.1% and 72.4% respectively.

With a validated dataset in hand, we use it to undertake an extended replication of Landes and Posner (2009). First, we do our best to replicate the original paper and, despite some problems in replicating the original dataset, we could replicate significance as well as the direction of the most important coefficients. We extend the results and probe their robustness using multi-way clustering, grouping, and additional covariates. Finally, we show that the results partially hold when using our machine-predicted ideological labels as the outcome.

This paper contributes to the emerging literature applying data science techniques to empirical legal research questions. We review some of that literature in Section 2. After that, in section 3 we describe the supervised learning task to predict ideological labels in circuit court decisions. Next, section 4 reports the results of our replication study. section 5 concludes.

3.2 Literature

This research sits at the intersection of two fields. On one side, our paper is related to the research on judge ideology, which is focused on the positioning judges, mostly for the U.S. Supreme Court (e.g. Giles et al., 2001; Epstein and Jeffrey A. Segal, 2005; Epstein, Andrew D. Martin, et al., 2012; Johnson et al., 2011; Kassow et al., 2012; Andrew D. Martin and Quinn, 2001; Masood and Songer, 2013; Ginn et al., 2015; Sturm and C. H. Pritchett, 2006; Randazzo et al., 2010; Reid and Randazzo, 2016).

In general, the judge ideology literature has taken two main approaches. The first approach is to hand-coded cases by ideological direction. These include the Spaeth database for the Supreme Court and the Songer database for the Circuit Courts (Epstein, Andrew D. Martin, et al., 2012; Sturm and C. H. Pritchett, 2006; Andrew D. Martin and Quinn, 2001; Epstein and Jeffrey A. Segal, 2005; Giles et al., 2001, e.g.). The second approach is to use a latent factor model based on the voting behavior, to estimate a latent dimension for ideology based on judge agreement. This approach can identify median judges and the relative judge positioning on a scale over time (Andrew D. Martin and Quinn, 2002).

The advantage of the first approach is that the scale is interpretable, exists on the case level, and relies on expert judgment. However, it is costly and there are errors in coding. The advantage of the second approach is that it is cheap to compute for all judges, but it is not directly interpretable and does not exist at the case level. It also requires that judges vote in panels.

Our approach is something of a compromise, as we can form predictions for all cases and judges cheaply. It requires at least some hand-coding, but then can be applied to all cases. Methodologically, it is different because it uses the directly interpretable ideological labels of the hand-coded database. It does not assume a latent factor model, like Andrew D. Martin and Quinn (2002). It also does not rely on contrasting votes of judges in a panel. This is relevant in our context because the large majority of decisions on the Appellate Courts do not have dissents. Voting behavior is not necessary, only some hand labels and the the original opinion text.

The second literature to which we contribute is that on using texts as data for social science research. In particular, to produce measures of ideology or partisanship. In law, an old study in this vein is Jeffrey A. Segal and Cover (1989), who use texts from newspaper editorials as a proxy for the ideology of newly appointed Supreme Court judges. More recently, popular methods in political science for scoring ideology in text include Wordscores (Laver et al., 2003), Wordfish (Slapin and Proksch, 2008), and Wordshoal (Lauderdale and Herzog, 2016). These tools use statistical differences in word frequencies by topic. They are most useful for text corpora for which differences in ideology come through in different words. As opinions of (lower) judicial courts are constrained in their (permitted) wording opinion texts may only satisfy that criterion in a very limited fashion.

In the legal domain, our paper is most closely related to literature predicting case type (Undavia et al., 2018; Sulea et al., 2017; Boella et al., 2011) as well as that concerned with political dimensions in judicial texts (see for example Ash and Daniel L. Chen, 2018; Ash, Daniel L. Chen, and W. Lu, 2018). The three papers closest to ours, in goal as well as methodological approach, are by Lauderdale and Clark (2014), Aletras et al. (2016), and Cao et al. (2018). In Lauderdale and Clark (2014), the authors use an LDA model to estimate how different issues at stake in cases are related to Supreme Court judges' voting behaviour. The paper by Aletras et al. (2016) looks at decision direction of the European Court of Human Rights (ECHR) in regard to the violation of specific articles. The third paper, Cao et al. (2018) separates opinion texts into ideological and fact-driven parts and look at how well these different paragraphs predict case directionality. However, none of the approaches in those three papers are viable for our goal or dataset. Lauderdale and Clark (2014) use the underlying text but their focus on votes means that the approach is not applicable. In the case of Aletras et al. (2016), in a modeling perspective the approach is similar. However, their results rely on very clean data resulting in very homogeneous directionality criterion. As a consequence, it is more than a simple question of transferring their results. Last, the paper by Cao et al. (2018) does look at ideological directionality. The focus on paragraphs, however, means that an additional labeling effort is needed while we seek to minimize the costs of classification.

To recap, our paper contributes in the technical literature to the understand how to best implement a machine learning approach in the domain of judicial opinions. We aim to decrease labeling cost and increase scalability and reproduciblity compared to the hand-labeling approach while at the same time improving explainability relative to the

latent modeling approach.

3.3 Supervised Classification

This section focuses on the classification algorithm which can reliably predict the political ideology of Circuit Court judges' written opinions. After training the algorithm on existing ideology labels, it can predict labels for unseen opinions.

The beginning of this section provides information about the data necessary for classification. What follows is a detailed description of how the classifier is trained. Finally, the classification performance is evaluated.

Data

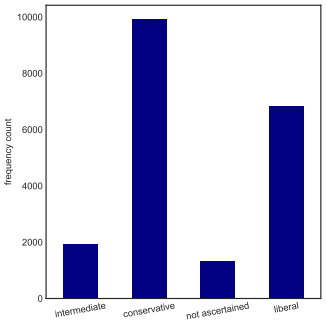
Broadly speaking, a supervised machine learning classifier maps an input to output. This section enumerates the datasets used for the inputs and outputs in our context. For our classification problem, we use the hand-coded ideology labels for these cases, provided by the Songer Project (Songer, 1993), as output. As input, we use the U.S. Circuit Court judges' written opinions.

Songer Data on Decision Direction

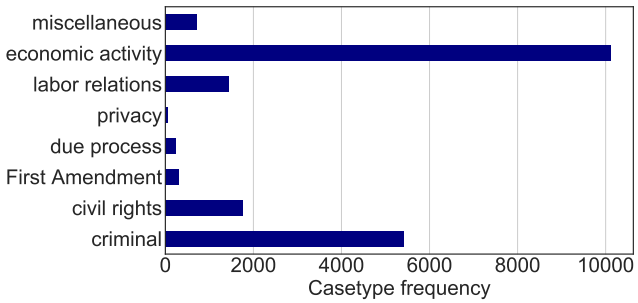
The output or label of our classifier is the ideological direction of the opinion. As the number of Circuit Court judges' opinions was comprised of over 300 thousand at the time of the database's inception, the Songer Project has annotated political ideology labels for only a small sample of opinions. These equal less than 2.6% of the total published opinions available. The cases were decided between 1925 and 2002 and the database contains a total of 20,355 cases. Overall, four directionality codes are available: "liberal", "conservative", "mixed" and "not ascertained". While "mixed" refers to the opinion of the case being of unclear directionality, "not ascertained" signals that the coders were unable to assign a label according to the codebook's instructions. Please note, that directionality is defined for each particular case type, with "conservative" and "liberal" being exactly opposite outcomes. Figure 3.1a shows the distribution of labels for the complete data-set. The categories "conservative" and "liberal" dominate, whereas the other two categories are underrepresented.

The Songer coders assigned the directionality of a case according to specific rules within case type. The case type of an opinion identifies the nature of the conflict between the litigants. Over 220 case type categories are organized into eight major categories: criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity and regulation, as well as miscellaneous. Figure 3.1b shows the distribution of the eight major categories for our data-set. "Civil rights" and "economic activity and regulation" are the two case types most frequent in the data.

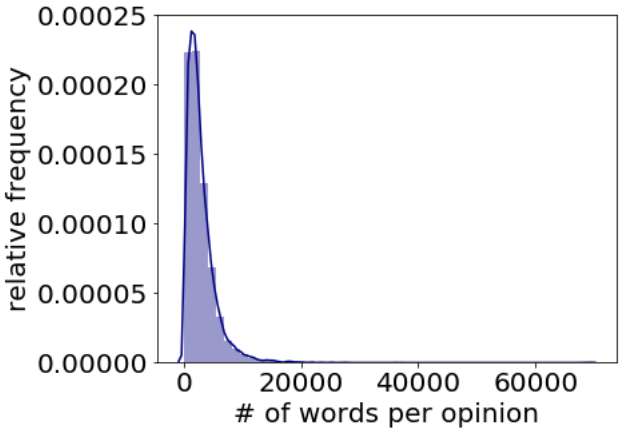
Landes and Posner (2009) mention in their paper that they applied substantial corrections to the raw Songer data, but those are not laid out in sufficient detail to



(a) Distribution of Political Ideology Labels



(b) Distribution of Case Types



(c) Distribution of Opinion Length (in Words)

Figure 3.1: Summary Statistics

reproduce. We approached the authors with the request to provide us with their version of the data-set. Unfortunately, they were not able to provide it yet.

Judicial Opinion Corpus

We matched the Songer data-set with the Lexis data-set, containing the full opinion text. With this approach, we could match 20,052 opinion texts to the 20,355 entries that the Songer database is comprised of. Regarding the non-matchable cases, there is no clear pattern visible as these cases span nearly the complete time period as well as nearly all circuits. The distribution across time and circuits does not reveal any peculiarities either.

In terms of the matching itself, we subsetted the data according to the different circuits. That was only done for speed, as matching is a linear searching process which has to be repeated for each query. The actual matching was then done on either federal

reporter citation or docket number. First, we tried to match via the normalized Lexis id, i.e. the Federal Reporter citation, if the opinion spanned more than one page in the Federal Reporter (to avoid confusion with other opinions). If such a match was not possible, we matched via the circuit court and the docket number. The reason why we preferred the federal reporter citation over the docket number is that the Songer database uses only encoded docket numbers. While they should be systematically, encoding errors often result in decoding being little more than guess work. In the case of the federal reporter citation, errors were less prone.

Figure 3.1c shows the distribution of opinions' word counts in our dataset. The shortest opinion consists of one word, the longest of 69,320 words. The average opinion consists of 2,809 words. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single newline character. Special characters such as "newline"-characters and roman numbers were removed.

If a potential heading was found within the text, we excluded it. The reason being that such a heading would potentially include biasing information such as judge names. It is especially important to exclude those, as the model could focus on judge names as a proxy for the directionality as most cases were decided without dissent. This is an issue in our empirical context because we would like to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions trying to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part which by its nature goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to an overall worse performance. If we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as dissent. We excluded those entries and did not use them as input.

Model

This section describes how we deploy a supervised learning approach to predict the ideological direction of decisions from the association opinion text.

Our approach, outlined by Figure 3.2, is quite uncommon in the literature of classifying a legal text's ideology. More traditional approaches, mainly used for ideology detection in political speeches, include word scores, word fish, or word shoal models. These approaches are either dictionary-based or require a reference text to which all other instances are compared. Our approach, by contrast, does not require one reference text to be selected and deploys more sophisticated selection mechanisms than naive word counts.

One characteristic of machine learning approaches is their exploratory nature. We, too, test multiple combinations of data-subsets, feature sets, models, and evaluation methods to find the best performing one. The instances to test are either selected by theoretical considerations, such as choosing only judicial quotations as predictive features; Or they are chosen based on popularity, such as choosing support vector machines because

they are known for their excellent performance on a broad range of NLP classification tasks.

All calculations were performed on the Max Planck Computing and Data Facility's high-performance cluster Draco, using one node of the type Broadwell with up to 40 CPUs and 256GB memory. Moreover, each step relying on randomness was initialized with a pseudo-random seed for replicability. Our code most heavily draws upon functionalities provided by the python package sci-kit learn (Fabian Pedregosa et al., 2011).

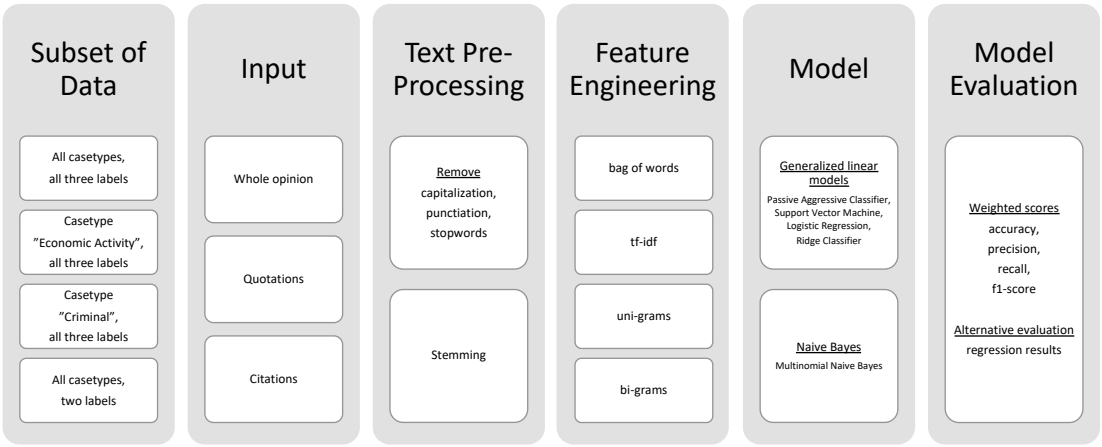


Figure 3.2: Construction of the methodological approach

Subset of Data. In order to see how different categories or a differing number of labels affects a prediction, we constructed different subsets of the data for analysis. The four subsets constructed from the original data and used for this analysis are listed in the first column of Figure 3.2. A naive approach predicts political ideology labels regardless of case type. However, the naive approach ignores the fact that directionality in the Songer data is assigned dependent on case type according to explicit rules differing for each case type. Subsetting the data by case type factors in this aspect of the coding scheme.

However, as Figure 3.1b shows, the dataset is heavily imbalanced in favor of the case types "economic activity" and "criminal". As the remaining case types are only marginally represented, we restrict the subset two these two case types, as only for them enough labeled observations are available to train the classifier.

Moreover, not only case type but also the labels are imbalanced. As Figure 3.1a shows, there is only a limited amount of observations available for the political ideology labels "not ascertained" and "mixed". We therefore derive two additional subsets. The subset "two labels‘" only includes the labels "conservative" and "liberal" as those two are not only the most frequent ones, but also those we are most interested in. Especially, if the remaining two labels ("not ascertained" and "mixed") are either considered to be noise or to be wrongly classified, this subset should improve the classifier's performance.

In particular, the exclusion of the label "not ascertained" is likely to not be problematic in any case: The number of cases labeled such are relatively few when compared to the other three labels. Moreover, the codebook shows that this label may be used in any case where it was not possible to assign one of the other three labels. This may either be due to the fact that the case truly fits into no other category or merely due to a lack in inter-coder agreement. However, past results show that such a sparsely represented, miscellaneous category decreases classification performance. For this reason, the final subset excludes this category altogether.

Input. We experimented with four different representations of the input. The most straightforward approach is to feed the complete pre-processed opinion text into the model. After screening a sample of randomly drawn opinions and cross-referencing them according to the labeling instructions from the codebook, we identified two additional representations.

First, we separately extracted the citations from the cases. The topic as well as the political directionality of a case might be captured already by citations. Citation networks, for example used by the Supreme Court Mapping Project, are one example using this reasoning (Chandler, 2005; Ash, Daniel L. Chen, and W. Lu, 2018).[2]

Second, we extracted quotations from the text to serve as input. Many quotations immediately preceded citations. It is in the nature of a quotation that it represents the most relevant aspects to a matter at hand. As judges quote legal concepts from statutes and precedents relevant to the matter discussed, quotations, in turn, may be associated with either a "conservative" or a "liberal" leaning of the opinion.

The advantage of the whole opinion text as input is that no information is lost. Its downside, however, is that it may include more noise than only citations or quotations.

Text Pre-Processing. For any data subset, the raw text needs to be pre-processed. We applied the prevalent practice of removing capitalization, punctuation as well as stopwords. Furthermore, we reduced the words to their word stem, base or root form (stemming).

Feature Engineering. The pre-processed text was tokenized, and the tokens were then used to form lists of n-grams (phrases) up to length three. N-grams extract information from text through local word order (Suen, 1979; Sidorov et al., 2014). In the next step, these tokens were mapped to a numerical representation. We computed counts and frequencies over n-grams. The second specification is to weight the counts (tf) by inverse document frequency (idf), which up-weights relatively rare words that could be more informative of topic or ideology.

Apart from converting opinion texts to vectors, we included the year the case was decided, the circuit at which the case was heard as well as the case type as assigned by the authors of the Songer database to the feature set, as well. Via grid-search, we established

[2] see SCOTUS Mapper Library by the University of Baltimore.

which input and pre-processing combinations worked best, especially regarding single words versus n-grams.

Model. After vectorization, the next step is the actual classification of the text input,[3] listed in the second last column of Figure 3.2. In general, the classifiers may be grouped into two families, with the first being statistical methods. The advantages of this family are high explainability, that it is being well-researched, and well understood (Ribeiro et al., 2016a). The second family is that of deep learning algorithms, mostly comprised of some form of neuronal network architecture. In common NLP tasks, these algorithms outperform traditional algorithms (Kim, 2014; Vaswani et al., 2017). However, a downside to these models is that feature introspection, as well as explainability, is difficult. While there are attempts to develop methods for feature introspection, such as Shrestha et al. (2017) or Ribeiro et al. (2016a), results so far are preliminary. Consequently, we focus on well-researched statistical classifiers, maximizing the explainability of the results. The classifier, we deploy are a passive aggressive classifier (Crammer et al., 2006), a logistic regression (M. Schmidt et al., 2017), a ridge classifier (Rifkin and Lippert, 2007), as well as a support vector machine with stochastic gradient descent (SGD) learning (Zhang, 2004). All models are trained on a stratified train-test split with respect to case type.

Model Evaluation. For model evaluation, we use standard performance metrics for machine learning, namely accuracy, precision, recall and f1-score (last column of Figure 3.2).[4] The f1-score is the harmonic mean of precision and recall. As compared to accuracy for example, it is more stable with respect to unbalanced data-sets like ours. Furthermore, in the context of this paper we consider precision as more important as recall, because our dataset contains much less liberal than conservative cases. Thereby, we consider it as more important to actually find theses few liberals and risk to classify some conservatives as liberal.

As all performance measures are 5-fold-cross-validated, the scores reported are weighted averages. As the label space per category is heavily imbalanced in the validation set, accuracy has to be interpreted with care, and therefore the best performing classifier is selected by referring to the weighted f1-score. In our case, an additional model evaluation is the use of the predictions in the replication analysis below.

Evaluation of Results

In the following, we provide in-depth analysis across the different classification models introduced by Figure 3.2.

[3]The classifiers are implemented with the python package sci-kit learn and fall into the category of supervised learning.

[4]While in traditional statistics measures such as the p-value are more prevalent, that measure is not appropriate in machine learning because we are trying to form accurate test-set predictions rather than to test for treatment effects. Moreover, the features in machine learning are often very highly correlated, so the estimated coefficients for them are difficult to tease apart.

Performance Metrics. Appendix C.3 depicts the performance metrics f1-score, accuracy, precision and recall for all models tested. Figure C.1 shows that the scores depend more heavily on the subset-input-combination than on the specific classifier used.

Based on this observation, we select four models to analyze and compare in detail. Figure 3.3 depicts the model for each of the four subsets tested which reaches the highest f1-score. We report the accuracy, precision, recall, and f1-score respectively (coded by color, see legend). Each of the four groups of bars refers to a different subset of the data, for which we explored different modeling approaches. The top row looks only at the liberal and conservative votes, dropping the "other" category. Second, we classify the full dataset with all three categories. Third, we limit the dataset to criminal cases. In the bottom row, we limit the dataset to economic cases.

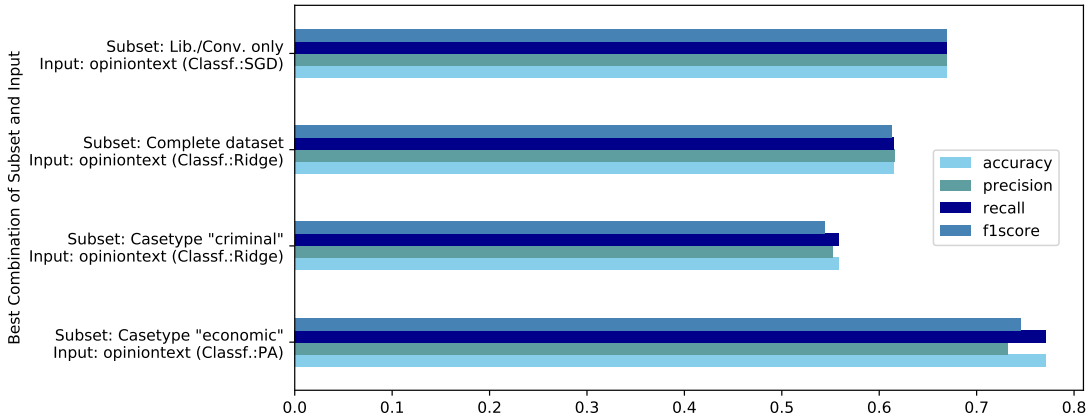


Figure 3.3: Best performing combinations by subset

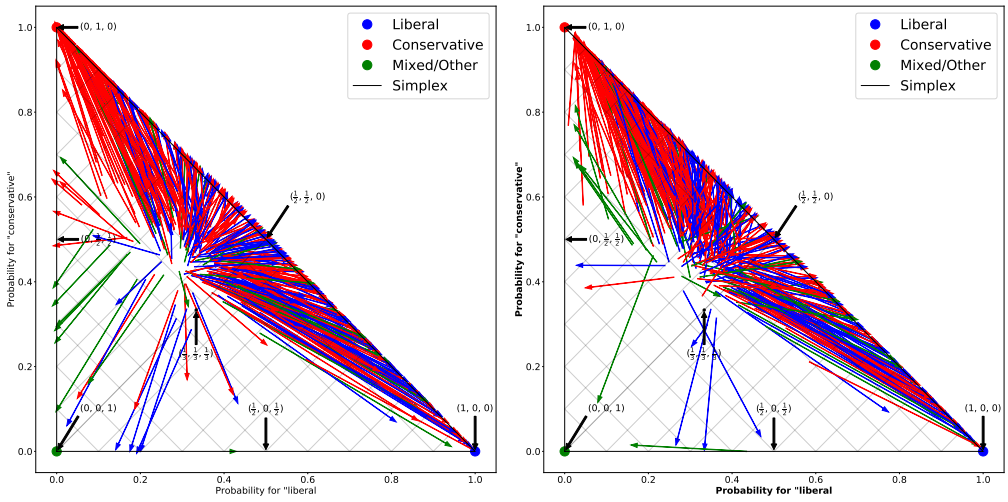
On the y-axis, we indicate a feature that all four models have in common: they perform best on the input opiniontext, rather than on citations or quotations. While additional calibration and tweaking of the model parameters would improve the performance of the classifier using either citations or quotations as input, the result is consistently outperformed when using the complete opiniontext as input. This observation contrasts with the idea that citations or quotations would summarize the information in a meaningful way. However, instead of subtracting what was assumed to be noise, it seems that these input variations subtract important information. As mentioned, the four subsets differ with respect to the subset of cases. Comparing the subsets concerning label, we differentiate between two or three label classification. The subset displayed at the top of Figure 3.3 takes two labels into account. A random guess, assuming a uniform distribution of labels, should yield an accuracy of approximately 0.5. The model reaches an accuracy of 67.04%, lying clearly above this threshold.

The second group of statistics are from the three-labels model. How much performance do we gain when predicting two instead of three labels? The two models at the top of Figure 3.3 show – only these two take all case types into account – an increase in accuracy from 62.00% to 67.04%. We believe that this increase in performance may offset the loss

of information by excluding the "mixed/other" label as less than 1/7 of all cases fall into this category. This opinion is shared by other authors, as well: Most studies drawing upon the Songer/Auburn database exclude the "mixed/other" cases. However, for the sake of thoroughness we undertake the calibration presented in the following section for both the two and three label subset.

In the third and fourth groups of performance metrics, we show the three-label model but subset on case type. Interestingly, performance depends strongly on the case type. As mentioned in subsection 3.3, directionality is defined within case type while the number and quality of rules are quite distinct. Additionally, as Figure 3.1 shows, case type is heavily imbalanced in favor of economic rather than criminal. These two facts help to explain why the subset criminal only reaches an accuracy of 55.80% and by contrast, why the subset economic achieves an accuracy of 77.10%. However, in order to increase generalizability, we instead opt to focus on classifiers trained on data containing all case types as some results from e.g. the case type "economic" may carry over to the case type "criminal".

Probability Calibration. In the following, we analyze our classifiers' calibration: Predicting a judicial opinion to either be conservative or liberal, we not only want to know the label but how confident the classifier is in assigning one particular label versus the other. In order to boost calibration, the classifiers were re-calibrated using either a sigmoid or an isotonic calibration function. The sigmoid function rests on a parametric approach based on Platt (1999)'s sigmoid model. The non-parametric isotonic variant is based on an isotonic regression.



(a) Ridge (Isotonic Calibration) (b) SGD (Sigmoid Calibration)

Figure 3.4: Drift-plots showing the Change of Predicted Probabilities after Calibration

Figure 3.4 depicts the Ridge and SGD classifier respectively. For both classifiers,

the calibration methods were applied for visualization purposes.[5] The three corners of Figure 3.4 correspond to the three classes: conservative, liberal, and mixed/other. Arrows point from the probability vectors predicted by an uncalibrated classifier to the probability vectors predicted by the same classifier after calibration. For clarity of presentation, only each fiftieth data point from the test set is depicted.

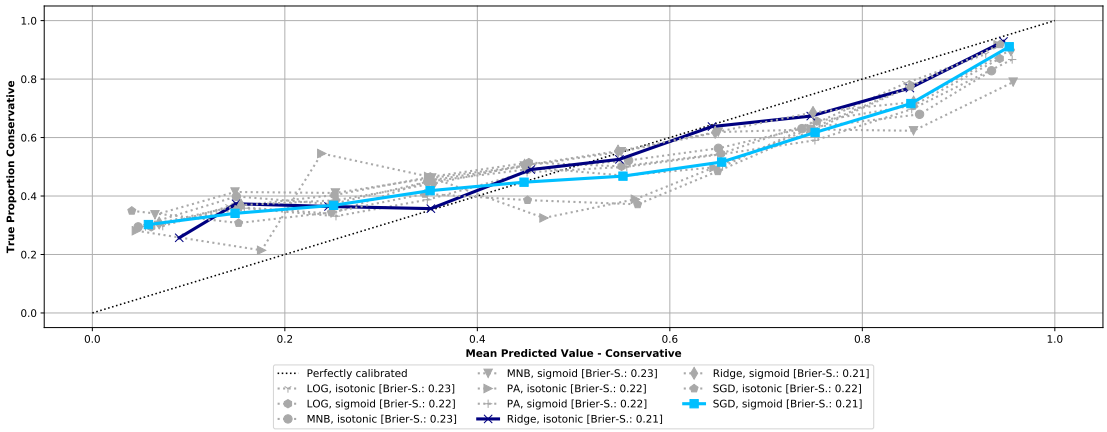
Figure 3.4 shows that calibration results in both classifiers shifting from under-confident to over-confident predictions. This can can be seen as the mass of predicted points moves away from the center of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ towards the edges. This means that the classifier is likely to categorize similar cases very differently as the predicted label is further away from the decision boundary for all cases. On the other hand, it also means that the classifier gets more confident about cases which are hard to classify – that is, the position of which is properly close to the decision boundary. We accept this change however, as the absolute accuracy as well as the f1-score increases, although there may be additional error for boundary cases.

While the two classifiers do not majorly differ in their confidence, they do differ in the error rate of assigning the label "liberal" to liberal cases. If one looks at the blue arrows, which depict cases for which the true label is "liberal", one can see that for the Ridge classifier (left panel) the mass of the blue arrows falls into the simplex spanned by the corner points $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$,$(1, 0, 0)$. Every arrow point found within this simplex is classified as "liberal". Consequently, as the mass of blue arrows falls into that area, the majority of them is categorized correctly. In contrast, for the SGD classifier (right panel) a lower amount of the blue arrows falls into that area, meaning that the misclassification rate for "liberal" is higher. This means the precision for liberal is lower for SGD compared to the Ridge classifier. On the other hand, the inverse is true for the recall. As the original dataset features fewer liberal cases than conservative, on balance we might prefer to mis-classify conservative cases as liberal instead of liberal ones as conservative. At this point, this speaks in favour of the Ridge classifier vs. the SGD classifier.

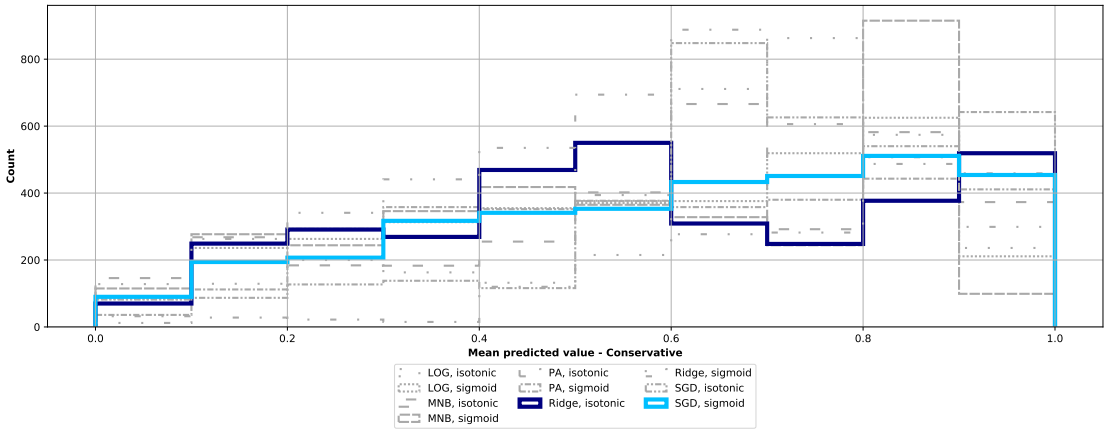
When looking at the "mixed/other" cases, we can see that the Ridge classifier classifies the majority of them correctly. However, that seems to come at the expense of misclassifying a disproportionally high amount of liberal cases. For the reasons stated above, we consequently exclude the "mixed/other" label to gain performance in predicting only the labels "conservative" and "liberal".

Figure 3.5a provides another visualization to assess how well the probabilistic predictions of different classifiers are calibrated: It displays reliability curves which show the correct proportion of conservative cases (vertical axis) against the bins of predicted probabilities that a case is conservative (horizontal axis). The closer the reliability curve is to the 45-degree line, the better is the classification model's performance in terms of reproducing the original distribution. The Ridge classifier with isotonic calibration, as well as the SGD classifier with sigmoid calibration are highlighted in shades of blue.

[5] Probability calibration was performed on data not used for model fitting. To this end, the training set consisting of 80% of the Songer data was cut in thirds and the model was then trained with 3-fold cross-validation. During this, 2/3 of the data were used for training and 1/3 was used for calibration. For each classifier, the calibration algorithm yielding the best results was chosen.



(a) Reliability curve



(b) Distribution Diagram

Figure 3.5: Reliability curves and Distribution Diagram

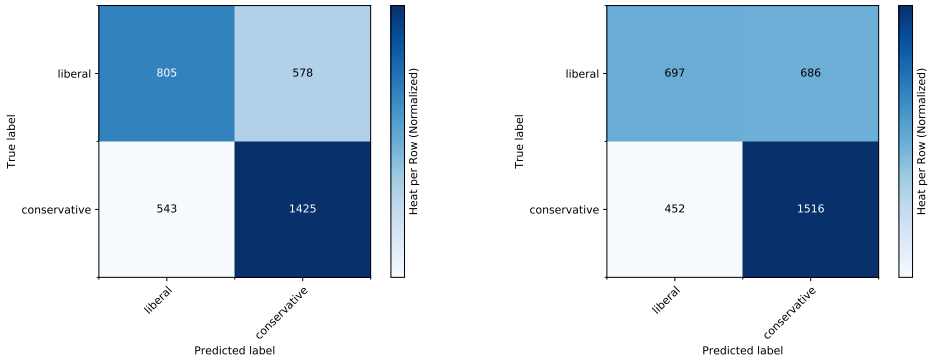
Consider the Ridge classifier: For all cases which it predicts to be conservative with a 20% probability, about 40% are actually conservative. In other words, it underestimates conservativeness. However, for cases close to the hyperplane (0.5 probability for either directionality), the classifier approximates the directionality distribution very well.[6] Finally, at around 70% likelihood, the classifier begins to overestimate the number of conservative cases.

Alongside Figure 3.5a, Figure 3.5b shows that despite calibrating the classifiers, a significant part of the predicted directionality's mass lies close to the decision boundary of 0.5. This, in turn, means that the classifiers have to be relatively precise close to the decision boundary and be able to shift away mass from the decision boundary. Figure 3.5b

[6]This is an important aspect as the Ridge classifier is similar to a support vector machine in that it uses the instances closest to the hyperplane for the separation of the data points.

shows that the two classifiers most successful in this are the ridge classifier, calibrated with the isotonic algorithm, and the SGD support vector machine, calibrated with the sigmoid algorithm.

Heatmaps. In the previous paragraph, we conclude that a two label classifier for all case types will be the basis for predicting political ideology labels. In terms of performance metrics, the SGD classifier reaches the highest f1-score. However, the decision for the final model should not just take the f1-score but rather the types of errors that the classifier makes into account, as well. Therefore, Figure 3.6 plots normalized[7] confusion matrices for those two models deploying the best f1-score: The Ridge as well as the SGD classifier.



(a) Ridge classifier (b) SGD classifier

Figure 3.6: Confusion Matrices for the classifiers SGD and Ridge

As mentioned in subsection 3.3, we consider it as crucial to correctly predict as many liberal cases as possible, even if some conservative cases are wrongly predicted as liberal. Figure 3.6b shows that as far as liberal cases are concerned, the SGD classifier predicts 697 cases correctly as liberal but almost as many cases (686) wrongly as conservative. The Ridge classifier displayed by Figure 3.6a, by contrast, predicts 805 liberal cases correctly as liberal and only 578 liberal cases wrongly as conservative.

Best classifier. Based on performance metrics, heat-maps, and calibration results, we can select the classifier most suited for the task set out in this paper. The f1-score – our preferred performance metric – peaks for the Ridge-Classifier, calibrated with an isotonic function as well as for the SGD-classifier, calibrated with a sigmoid function. The second performance metric we consider as critical is precision, for which the Ridge classifier shows better results than SGD. In the same vein, the reliability curves show that Ridge is closer to the 45-degree line than SGD, which makes the former preferable. The only aspect where the SGD support vector machine slightly outperforms the Ridge

[7] normalized heat is calculated by dividing each value by the row mean

classifier is in terms of mass, as shown in Figure 3.5b. However, overall, the difference in this regard is negligible. Given this reasoning, we chose the Ridge-classifier calibrated with the isotonic algorithm as model to perform out of sample predictions.[8]

Analysis

This section analyzes and interprets the predictions of the best two-label classifier. We look at predictions over time and by judge. We also interpret the model by examining predictive features.

Prediction of the Time Series in Decision Direction. Landes and Posner (2009) point out that the accuracy of the original Songer data is susceptible to the year in which a judge decided a case. Coders had more trouble coding older cases as compared to newer ones. We would like to see if this is reflected in differential performance of our classifier over time.

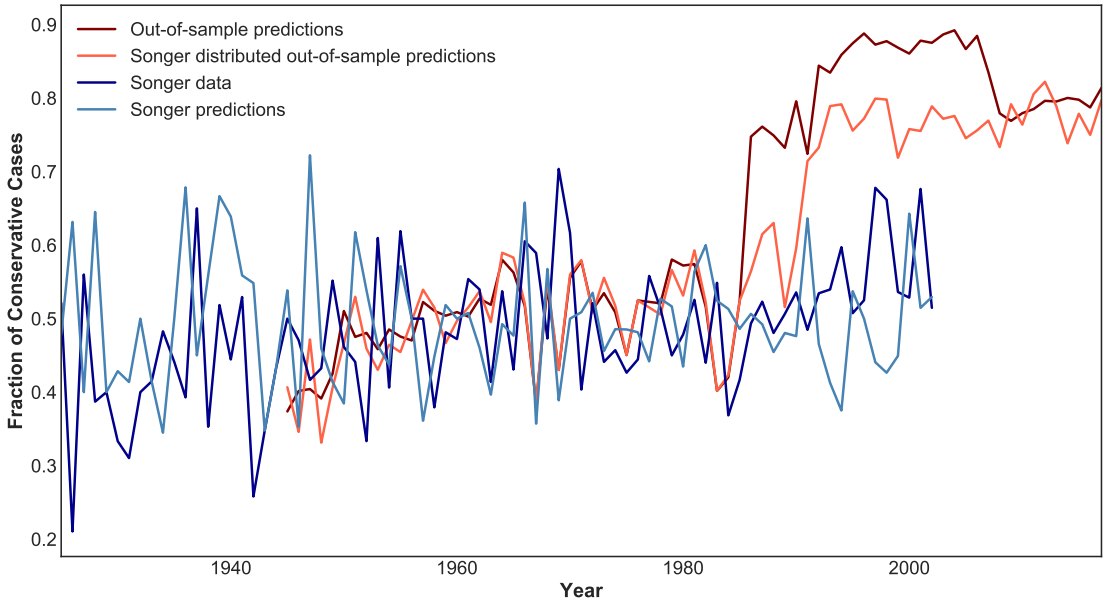


Figure 3.7: Fraction of conservative and liberal cases, each calculated for actual as well as predicted case directionality, plotted by year

Figure 3.7 shows the fraction of conservative and liberal cases by year for all circuits.[9] We include out-of-sample data which is made up of scraped lexis data without the cases

[8]The final specifications of the classifier are as follows: We preprocess the text by excluding all stop words as well as punctuation. Following that, a lemmatizer is applied. This input transformed into bigrams and then fed to a tfi-vectorizer. That vectorizer calculates the distance based on the "l2"-norm. It also makes uses the three additional features of year, circuit and case type. The regularization strength parameter α for the Ridge classifier is 2.0

[9]The cases categorized as "mixed" or "other" are excluded.

already within the Songer dataset. The original scraped dataset holds more than 1 million cases. As our classifier uses the year of the case, the circuit, and the case type as laid out by Songer[10] these features have to be available for all out-of-sample cases, as well. Especially the last one constrains the lexis dataset because case type was only available for cases of the years 1930 and later. Consequently, Figure 3.7 shows out-of-sample predictions only for those years.[11]

Figure 3.7 shows that for the in-sample predictions on the test set of the Songer data (20% hold-out data), the predictions closely approximate the original labels. This is also reflected in the high correlation of 0.73 ($\alpha < 1\%$). Especially for the years 1950 to 1980, the classifier performs very well. The out-of-sample predictions for that time period approximate the trend observed in the Songer data. Only for the years of 1980 on-wards, the out-of-sample data (red line) is predicted to be considerably more conservative.

This spread may be caused amongst others by the classification error. Another reason could be the sampling process used by Songer and his team to construct the database.[12] To test this presumption, we plot a subset of the lexis data constructed according to Songer's rules ("Songer-distributed out-of-sample", the orange line). Indeed, we find that the orange and red lines diverge after 1980, with the orange line being closer to the original Songer data. This illustrates that indeed the sampling process heavily influences the distribution of decision directionality: As soon as the total amount of cases increases by a significant amount, a spread appears.[13] As the absolute number of court cases increased over time (Casper and Posner, 1974), at least for cases after 1980 the Songer data may not be a good sample for the full set of cases. Consequently, the difference in out-of-sample predictions as compared to Songer predictions may simply stem from the fact that there is a structural shift in conservativeness (either in variation or trend) from 1980 onward which is not represented by the Songer sample.

Directed Votes per Judge. Next we zoom in on particular judges. We look at performance for the ten judges who cast most of the votes in the Songer dataset, analyzing performance in civil and criminal cases separately. Those judges who did not hear both civil and criminal cases were excluded. The horizontal axis of Figure 3.8 indicates the true proportion of conservative votes, while the vertical axis indicates the predicted proportion of conservative votes. Each point indicates these statistics for a single judge. If a judge's predicted behavior is the same as the truth, then his/her data point would lie on the dotted 45-degree line. Figure 3.8 shows that for civil cases, predicted and actual fractions

[10] We matched the lexis case types to the one laid out in the Songer database. However, the match is non-bijective. In order to get a reasonable good match, the subcategory case types of both, the Lexis data base as well as the Songer data base were used. This match is surjective with the Lexis subcategory case types as a base set. Then the matched Songer sub categories are aggregated to a Songer top category. Except for very few cases (< 1000), this aggregation is unequivocal.

[11] If one is willing to forgo the performance gain introduced by the case type feature (about 2.5% points in the current configuration), one can predict directionality for all lexis cases.

[12] For the original Songer database, at maximum 30 cases per year per circuit were sampled from all available cases after 1961. Before 1961, only 15 cases per year per circuit were selected.

[13] Where for the year 1945 only slightly more than 100 cases per year per circuit were coded with a usable

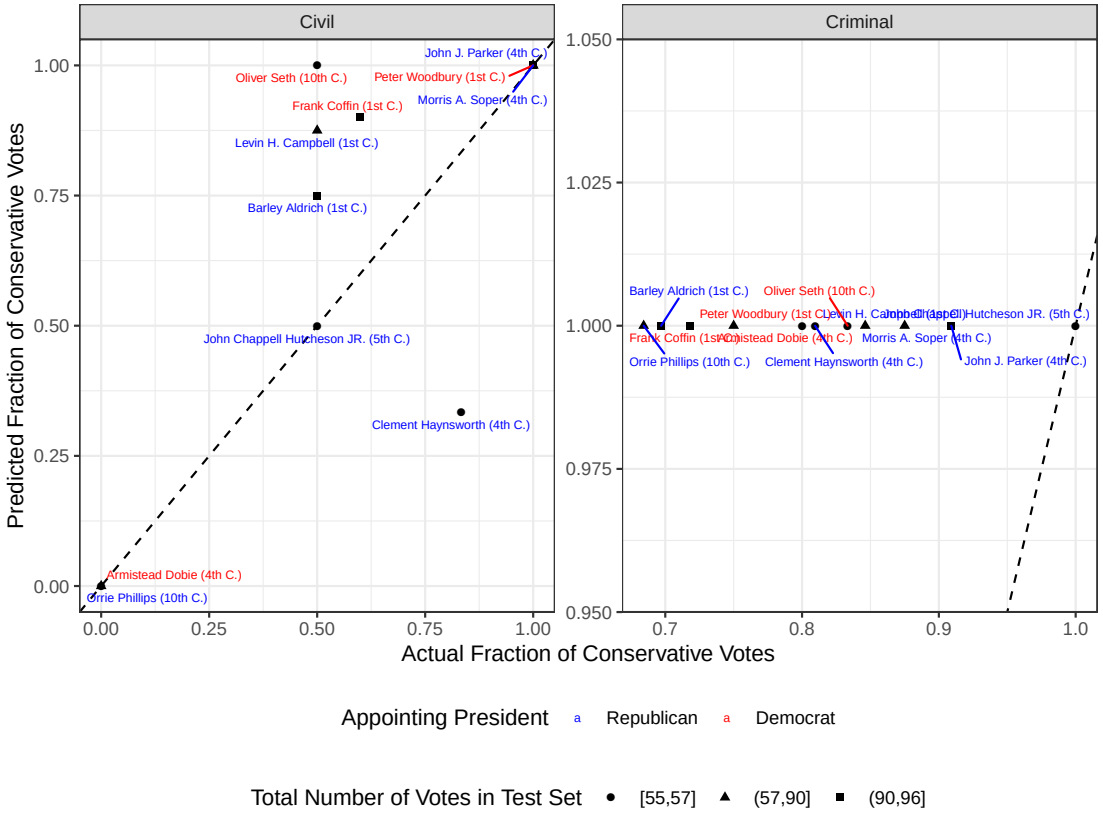


Figure 3.8: Fraction of Directed Votes per Judge - Comparison Actual Votes and Predicted Votes

are quite close. A χ^2 test shows that the distribution of predicted fractions is not statistically different from the distribution of actual fractions ($p^{\chi^2} > 0.1$). For case type criminal, however, the distributions of true and predicted fractions across judges are statistically different. The reason for this might be that the majority of criminal cases is labeled as conservative. Consequently, as the classifier uses the case type as feature it can increase performance on criminal cases by labeling it as conservative. In other words, the classifier tends to overpredict the number of conservative cases in criminal law.

Feature Inspection. To further understand the two-label classifier, we investigate the features that are most important in driving our predictions. For this purpose, let $feature$ be a feature, $value$ be a value it could take, and $label$ one of the ideological directions (conservative or liberal). We ranked the informativeness of each feature by the highest value of $P(feature = value|label = conservative)$ divided by $P(feature = value|label = liberal)$. Note that these are equivalent to coefficients from

case type in the out-of-sample dataset, for the year 2000 there are more than 2000 per year per circuit.

a Naive Bayes Classifier. The coefficients of the different features are represented by their standardized moments, meaning that normalization was performed by dividing through the standard deviation. This means that each coefficient is on the same scale and therefore comparable. The hyperplane separating "conservative" from "liberal" lies at 0, meaning a hypothetical case for which all the decision results would be zero falls into neither category. The higher the coefficient of a feature, the further away does a single feature move the case instance from the hyperplane when the feature is present within the case. Table 3.1 lists the most informative features used by our best performing classifier. Please note that the most informative features for the label "liberal" are constructed such that they are least informative for the label "conservative". The features are either opinion-text phrases, quotation phrases, or citations.

Table 3.1 shows that the coefficients differ vastly in absolute size across the three different input variations. This corroborates the results of the metric scores. Especially for citation as input, the range of the coefficients' values is very narrow, with -7.49 being the minimum and 10.16 being the maximum. Consequently, many features loading clearly on either the "liberal" or the "conservative" side are needed in order to have the case fall into a category. By contrast, the range of the coefficients' values for opiniontext is much wider, with a minimum of -57.67 and a maximum of 189.96. A case including the words "reverse remand" for example would be classified immediately as liberal. In essence, this means that features for the opiniontext or quotations as input are more informative than for the citations. The first column of Table 3.1a and Table 3.1b have the most predictive quotations. Quotations loading heavily on the label "conservative" are "knowingly" or "unique circumstances". The court quotes these phrases, i.e. they are singled out as relevant to the case at hand. Both phrases indicate a possible conviction. As the code book by the authors of the Songer database very often label a conviction as "conservative", this seems to be in line with the data provided. On the other side, the quotations for "liberal" are not as easily interpreted.

The second column of Table 3.1a displays those citations loading on the label "conservative". For the most heavily conservative citation, Humphrey v. Moore, the court limited the power of unions from infringing too far on employees of a company not part of the union. In Dandridge v. Williams, the court found that the state has some right to interpret how it puts into practice federal welfare laws. In consequence, Maryland was found not to be in violation of the anti-discrimination act. Another conservative example would be United States v. Robinson, in which the court strengthened the police powers for searches during lawful arrests under the fourth amendment. In comparison, in the second column of Table 3.1b features citations which the classifier finds to be indicative of a liberal case. The most indicative citation would be United States v. Taylor, a case in which the bar for conviction on charges of conspiracy was raised. Coppedge v. United States dealt with the fact that the sentenced petitioner had not received the plenary review of his conviction to which he is entitled and all his appeals against his conviction against this ground were dismissed. The Supreme Court reversed the decision to dismiss his appeal

and generally strengthened defendants rights in this regard. In the same vein, Green v. United States reversed the sentencing of the defendant under the Fifth Amendment as he was put in jeopardy twice for the same offense. Consequently, while absolute size of the coefficients for citations hint at only a limited quality for the overall classification into either “liberal” or “conservative”, the cases as such seem to fall into the right domain.

The last column shows the predictive phrases from the full opinion text. Features such as ”judgment affirm” or ”plaintiff appeal” are predictive of the conservative label. In line with those but not shown here are the features ”affirm judgment” and ”appeal dismiss” on place 11 and 14 respectively. This is in line with labeling rules as set out by the Songer team for criminal cases, where the coding rules state that affirming the decision against an appellant is to be coded as conservative. Conversely, within the most predictive features for “liberal” one can find ”reverse remand”, ”remand proceeding”, or ”reverse cased”, reflecting that predictive features seem to be driven by criminal cases.

Table 3.1: Best Predictive Features

(a) Best Predictive Features for Label “conservative”

	quotations (Ridge)		citation (Ridge)		opiniontext (Ridge)	
	coef	feature	coef	feature	coef	feature
1	-17.13	knowingly	-7.49	Humphrey_v_Moore	-57.67	motion new
2	-13.18	John_Doe	-7.43	Dandridge_v_Williams	-53.71	plaintiff argue
3	-11.97	unique_circumstances	-6.59	SEC_v_Chenery_Corp	-51.91	prior art
4	-11.47	X	-6.42	Co_v_Zenith_Radio_Corp	-50.86	appellant claim
5	-11.40	No	-6.19	Dalehite_v_United_States	-50.78	grant motion
6	-11.03	minor	-6.06	Brady_v_Maryland	-49.45	plaintiff appellant
7	-10.85	search	-5.60	United_States_v_Robinson	-48.85	plaintiff contend
8	-10.63	attractive_nuisance	-5.55	Mal_v_Riddell	-45.70	fiduciary duty
9	-10.09	may	-5.38	Port_Gardner_Investment_Co_v_U	-45.62	plaintiff appeal
10	-10.04	overhead	-5.25	Olim_v_Wakinekona	-44.01	judgment affirm

(b) Best Predictive Features for Label “liberal”

	quotations (Ridge)		citation (Ridge)		opiniontext (Ridge)	
	coef	feature	coef	feature	coef	feature
1	19.98	that_where_the_State_has_provided_an_opportuni...	10.16	Yes_v_United_States	189.96	reverse remand
2	19.86	Motion_for_Judgment	9.18	United_States_v_Taylor	133.90	remand proceeding
3	19.57	fairer_to_those_adversely_affected_by_a_bond_f...	9.11	...Inc_v_Commissioner	103.28	case remand
4	19.16	take_care	9.09	Townsend_v_Sain	98.70	remand district
5	18.30	urge_that_the_indictment_charged_the_maintenan...	8.88	United_States_v_Young	89.69	government argue
6	17.32	good_faith	8.43	Dennis_v_United_States	85.99	remand new
7	17.30	anything_of_value	8.21	Coppedge_v_United_States	84.05	proceeding consistent
8	16.76	crack a_little_bit_of_time_to_research_on_the_...	8.15	...Inc__v_United_States	75.33	consiStatent opinion
9	16.76	a_little_bit_of_time_to_research_on_the_backgr...	8.00	Green_v_United_States	74.29	new trial
10	15.49	clear_and_convincing	7.97	Brown_v_Board	60.13	reverse case

3.4 Replication and Robustness Checks

This section focuses on the replication aspect of Landes and Posner (2009). For comparison, all tables and figures that Landes and Posner (2009) produce with data of Circuit Courts are listed in Table C.1, section C.1. The most relevant tables for our purposes are Tables 11 and 13, as numbered in the original paper.

Summary Statistics. This paragraph compares our summary statistics listed in Table 3.2b to those by Landes and Posner (2009, p.803) listed in Table 3.2a. As can be seen, the statistics differ. We count a total of 56,602 cases; Landes and Posner (2009) count 55,041 cases. Furthermore, we count more opinions classified as "conservative" or "other" than Landes and Posner (2009) do.

One possible explanation for these diverging results is, that not all of corrections that Landes and Posner (2009) applied in the original paper were described in sufficient detail to reproduce. We were able to apply the corrections concerning political ideology (Landes and Posner, 2009, pp.830-831) but we were unable to apply judge-related corrections. Landes and Posner (2009) briefly mention judge-related corrections and refer to a website for a detailed description. This website however, is no longer available online.

Table 3.2: Court of Appeals Votes by Subject Matter and Ideology for 538 Court of Appeals Judges Only: 1925 - 2002

	Crim	Civ Rts	First	Due Proc	Priv	Labor	Econ	Misc	Total
Conservative	6823	2721	566	461	117	1351	9361	525	21925
Liberal	1876	1766	477	201	67	1922	9884	559	16752
Mixed	635	460	89	51	13	420	1775	22	3465
Other	5321	210	102	79	3	179	6047	958	12899
Total	14655	5157	1234	792	200	3872	27067	2064	55041

(a) Original by Landes and Posner, 2009

	Crim	Civ Rts	First	Due Proc	Priv	Labor	Econ	Misc	Total
Conservative	7217	2647	397	412	83	1397	11084	478	23715
Liberal	1911	1755	379	176	38	0	10375	596	15230
Mixed	613	473	86	48	9	423	1689	31	3372
Other	5652	212	40	24	3	2232	5177	945	14285
Total	15393	5087	902	660	133	4052	28325	2050	56602

(b) Replication

Regression. Next, we replicate the primary regression analysis of circuit court judges in Landes and Posner (2009), focusing only on the essential part of their analysis. For

Table 13, we replicate the regressions focusing on the fraction of conservative votes and only taking the period from 1925 to 2002 into account.[14]

Regarding the baseline regression, Landes and Posner (2009) specify their regression model as follows:

$$FrCon_{ij} = \beta_0 + \beta_1 X_i + w \tag{3.1}$$

where $FrCon_{ij}$ denotes the fraction of conservative votes, calculated as votes per judge over the sample period. X_i encompasses several judge characteristics such as the party of the appointing president, share of Republican senators at the time of nomination, year of appointment, gender, race[15], prior experience as a district judge, as well as judge circuit fixed effects[16] According to Landes and Posner (2009, p.810), their regressions are weighted either by the judge's total votes in civil cases or the total votes in criminal cases. Furthermore, Landes and Posner (2009) do not specify how they compute their standard errors, but we assume that they use heteroskedasticity-robust standard errors (treating each judge as an observation) and therefore use errors of that type for the replication.

Civil Cases

In Table 3.3, we provide our first replication table, dealing with civil cases only. Column (1) corresponds to Landes and Posner (2009) Table 13 column (6).[17] As in the original paper, we report the t-statistics, rather than standard errors or p-values, for all coefficients in parentheses. Landes and Posner (2009) do not specify how they computed standard errors for their regression Table 13, but we inferred that they used heteroskedasticity-robust errors.

The main research interest of Landes and Posner (2009) was whether judges follow their party affiliation in their decisions. They find a significant influence of being appointed by a Republican president (RepPres) on the fraction of conservative votes for civil cases (Table 3.3, column 1). Our result for civil cases (Table 3.3, column 2), is quite similar when compared to Landes and Posner's; in our data, being appointed by a Republican is associated with a positive and significant effect of voting conservatively in civil cases. The evidence for a relationship between party and ideology actually appears to be stronger in our replication than implied by the original study.

Apart from deploying heteroskedasticity-robust errors, we propose a model specification with multi-way clustering (non-nested) as recommended by Cameron et al. (2006). Based on the advice from Abadie (2018), we add two-way clustering by circuit and year.

[14] In turn, this means that we do not display results for the fraction of liberal votes, as displayed in columns (2) and (4) of Landes and Posner (2009) Table 13, nor do we report results for the period of 1960 to 2002 as reported in Table 14.

[15] Race is a dummy for Black = 1, 0 else

[16] The judge specific data was acquired from the Auburn database by Gary Zuk, Deborah J. Barrow and Gerard Gryski on http://www.songerproject.org and then matched to the Songer data by a judge identifier code.

[17] These are the columns with the "uncorrected" data. We only compare uncorrected data as Table 3.2 showed that we were not able to replicate even summary statistics for the corrected version.

Table 3.3: Regression Analysis of Court of Appeals Votes: 1925-2002, Civil Cases

	Dep. Variable: Fraction of Conservative Votes					
	true data				*predicted data*	
	Landes (2009)	replicated	multi.clus	vote	multi.clus.pred	vote.pred
	(1)	(2)	(3)	(4)	(5)	(6)
RepPres	0.035***	0.069*	0.069***	0.092***	0.032**	0.031
	(3.860)	(2.125)	(4.136)	(3.821)	(2.942)	(1.417)
SenRep	0.072	−0.017	−0.017	0.095	0.004	0.219
	(1.710)	(−0.090)	(−0.347)	(0.647)		(1.677)
YrAppt	0.0003	0.001	0.001	0.0003	0.001	0.001
	(0.790)	(0.665)	(1.237)	(0.202)	(0.796)	(0.431)
Gender	−0.006	0.015	0.015	−0.026	−0.0004	−0.058
	(0.260)	(0.344)	(0.318)	(−0.681)	(−0.011)	(−1.384)
Black	−0.028	−0.105	−0.105	0.007	−0.125	−0.001
	(1.180)	(−1.505)		(0.124)		(−0.023)
DistrictCourt	0.002	−0.004	−0.004	−0.002	−0.002	−0.0005
	(0.330)	(−1.455)	(−1.183)	(−1.712)	(−0.345)	(−0.417)
FracEcon	−0.090	−0.230	−0.230**	0.355**	−0.249	0.451***
	(1.640)	(−1.506)	(−2.690)	(2.774)	(−1.918)	(3.531)
FracMisc	−0.049	1.345*	1.345*	−0.920	1.464***	−0.324
	(0.350)	(2.442)	(2.107)	(−1.842)	(6.118)	(−0.673)
circuit FE	yes	yes	no	no	no	no
circuit-year FE	no	no	yes	yes	yes	yes
Observations	535	498	498	4169	498	4169
R^2	0.240	0.119	0.119	0.047	0.123	0.066

*p<0.05; **p<0.01; ***p<0.001.
Linear regression with heteroscedasticity robust standard errors.
Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge ;*DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: all regressions include 11 dummy circuit variables - circuits 1 to 11 with the D.C. court the omitted circuit variable.

This allows for correlation in the error term across judges within court over time, as well as across courts in the same year. Clustering leaves coefficients unchanged, and a comparison of columns (2) and (3) reveals that t-statistics only differ slightly as a result of the two-way clustering.[18]

While Landes and Posner (2009) grouped the data on judge-level, we additionally run the empirical analysis with data at the vote level. This specification allows us to control for case characteristics with circuit-year fixed effects. For informing on the effect of party affiliation on ideology, this is econometrically an important step. The underlying reason is that the number of Republican-appointed judges and the proportion of conservatively decided cases could be correlated over time due to unobserved confounding factors. We also binarize the dependent variable. It equals one for conservative decisions and zero for liberal decisions. The cases with belonging to the "mixed/undetermined" category are dropped. The vote level regression model includes circuit-year fixed effects, as well as clustered standard errors by judge and year. This specification successfully replicates the significant positive effect of a conservative appointing president (RepPres) on the fraction of conservative votes.

Model specifications (5) and (6) are estimated not only with hand-labeled but also with predicted data. The predictions on which estimation results of columns (5) and (6) are based, were generated with a calibrated Ridge classifier. These re-estimations serve as an alternative way to assess the performance of the classifier. The rationale behind this procedure is that generating labels is not the end-goal, but using these labels in an empirical model is. Therefore, even if the classifier cannot predict political ideology with an accuracy of 100 percent, its performance can be viewed as appropriate if the results of the empirical model do not change drastically when estimated with the classifier's predictions. As far as column (5) is concerned, using predicted instead of hand-labeled data does not change the results for coefficients RepPres. Estimating the vote level fixed effects model with predicted labels instead of hand-labeled (column 6) results in estimates for RepPres that are no longer statistically significant.

Criminal Cases

In Table 3.4, we provide our second replication table; it shows criminal cases only. Landes and Posner (2009) found a positive and significant influence of being appointed by a Republican president (RepPres) on the fraction of conservative votes. Our result for criminal cases is quite similar to Landes and Posner's, our coefficient being slightly larger. Furthermore, for criminal cases, Landes-Posner found a negative effect of appointment year. However, we do not find such an effect. They also report a negative impact of being black (Black) on crime conservatism, which we replicate. Applying the two-way clustering changes the t-statistics only slightly and thus leads to no change in significance-levels

[18] We provide regression results with errors clustered on the year of appointment, Circuit Court, and the party of appointing president in Table 3.3, column (3).

Table 3.4: Regression Analysis of Court of Appeals Votes: 1925-2002, Criminal Cases

	Dep. Variable: Fraction of Conservative Votes					
	true data				*predicted data*	
	Landes (2009)	replicated	multi.clus	vote	multi.clus.pred	vote.pred
	(1)	(2)	(3)	(4)	(5)	(6)
RepPres	0.056**	0.077***	0.077***	0.051**	0.038	0.005
	(4.220)	(3.634)	(3.811)	(3.022)	(1.734)	(0.829)
SenRep	−0.076	−0.151	−0.151	0.010	−0.020	0.078**
	(1.090)	(−1.399)		(0.141)	(−0.542)	(2.844)
YrAppt	−0.001***	−0.00001	−0.00001	−0.0003	0.001**	−0.001**
	(3.390)	(−0.023)	(−0.032)	(−0.601)	(2.876)	(−2.709)
Gender	−0.014	−0.019	−0.019	0.010	−0.023*	−0.012
	(0.710)	(−0.740)	(−0.876)	(0.545)	(−2.219)	(−1.750)
Black	−0.057*	−0.091*	−0.091	−0.081**	−0.020	−0.027
	(2.060)	(−1.814)	(−1.047)	(−2.717)	(−0.257)	(−1.697)
DistrictCourt	0.001	−0.001	−0.001	0.0003	−0.001	0.001
	(0.140)	(−0.817)	(−0.390)	(0.360)	(−0.346)	(1.917)
circuit FE	yes	yes	no	no	no	no
circuit-year FE	no	no	yes	yes	yes	yes
Observations	523	498	498	13543	498	13543
R^2	0.240	0.084	0.084	0.019	0.052	0.014

*p<0.05; **p<0.01; ***p<0.001.
Linear regression with heteroscedasticity robust standard errors.
Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge ;*DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: all regressions include 11 dummy circuit variables - circuits 1 to 11 with the D.C. court the omitted circuit variable.

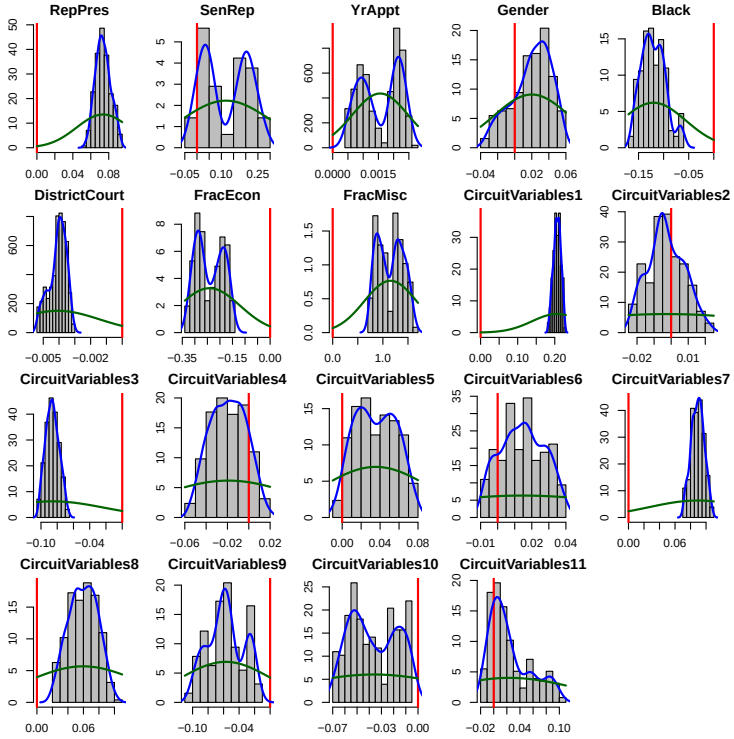
for the coefficient (RepPres). However, the coefficient for (Black) loses its significance upon application of the method. The fixed-effects multi-way clustering model on vote level data replicates the significant and positive effect of the party of the appointing president (RepPres) as well as of being black (Black) on the fraction of conservative votes. However, the multi-way error component model using predicted data could not reproduce the significance of the coefficient RepPres. Instead, being male turned to have a significant negative impact on criminal conservatism. Moreover, the fixed effects multi-way clustering model on vote level with predicted data could neither reproduce the significance for coefficient RepPres nor Black.

Extreme Bounds Analysis

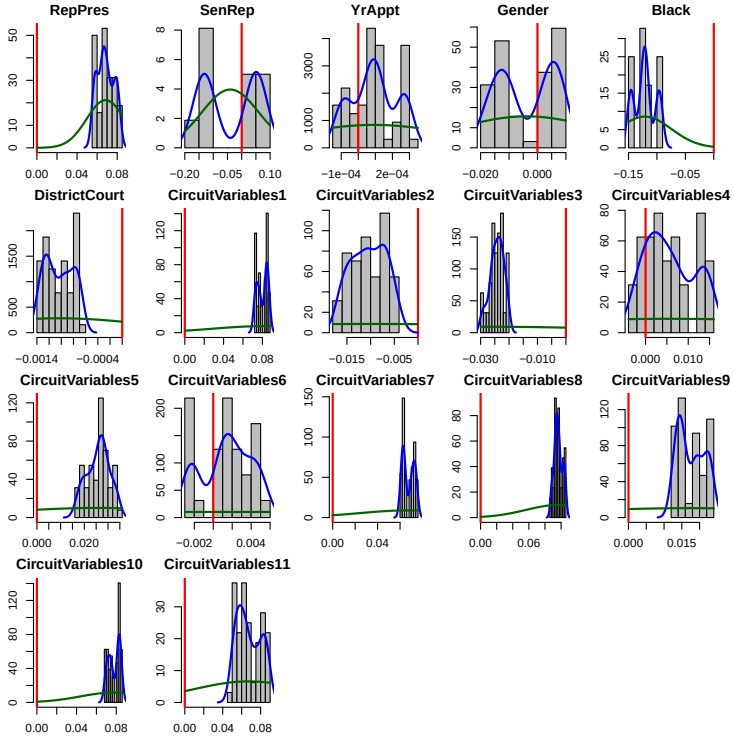
The extreme bounds analysis (EBA) is a sensitivity test that examines how robustly the dependent variable of a regression model is associated with a variety of possible determinants (Hlavac, 2016). We estimate an EBA, including all possible combinations of independent variables that Landes and Posner (2009) specified. To limit the influence of coefficient estimates with high multicollinearity, we follow the recommendations by Hlavac (2016) and specify the maximum acceptable variance inflation factor to be 7. Next, we increase the weights of those regression models that better fit the data – that is, by its likelihood ratio index according to McFadden (1973). Figure 3.9 shows histograms for each of the independent variables included in the model. The green curve displayed in each histogram is a density curve which approximates the coefficients' distribution with a normal distribution.

A positive coefficient indicates that holding all else equal, a higher value of the examined variable is associated with a higher fraction of conservative votes. On the other hand, if most of the area of the histogram's bins lies to the left of zero, higher values of the corresponding variable are associated with a lower fraction of conservative votes. For the civil cases, Figure 3.9a suggests that when the appointing president (RepPres) is Republican (rather than Democrat), when the judge was appointed in later years (YrAppt), as well as when the specific judge participated in a higher fraction of miscellaneous votes (FracMisc), a judge's fraction of conservative votes increases. Furthermore, circuits 1 and 7 are consistently associated with a higher fraction of conservative votes. Being black (Black), having served more years as a district judge (DistrictCourt), as well as an increasing fraction of economic votes (FracEcon), are associated with a lower fraction of conservative votes. Furthermore, circuits 3, 9, and 10 have a lower fraction of conservative votes. To conclude the visual inspection as well as the interpretation of the statistics, found in section C.5, the EBA for civil cases suggests that the variables RepPres, FracMisc and circuit 1 are very strongly associated with the dependent variable.

For criminal cases, Figure 3.9b shows that being appointed by a Republican (rather than Democrat) president (RepPres) is consistently associated with a higher fraction of conservative votes for all regression models estimated. Furthermore, circuits 1, 5, 7, 8, 9, 10, and 11 are associated with a higher fraction of conservative votes. By contrast, being black (Black) as well as having served more years as a district court judge (DistrictCourt)



(a) Civil Cases



(b) Criminal Cases

Figure 3.9: Histograms Extreme Bounds Analysis, for Civil and Criminal Cases

decrease the fraction of conservative votes. Furthermore, circuits 2 and 3 are associated with a lower fraction of conservative votes. The outlined findings, in conjunction with those laid out in Appendix C.5, EBA results for criminal cases suggest that the variables Pres, Black as well as circuit 8 and 10 are robustly associated with the fraction of conservative votes.

3.5 Conclusion and Outlook

This paper had two main goals. First, we aimed to replicate the analysis on Circuit Courts proposed by Landes and Posner (2009), and to add multiple robustness checks to assess the validity of the regression model initially specified. Second, we showed an approach for extending the dataset used in the original study via machine learning, especially in regards to the input used for any future algorithm. As far as the replication of the empirical analysis of Landes and Posner (2009) is concerned, we were able to reproduce the most critical findings. When we include additional robustness checks, we found – corresponding to the initial results by Landes and Posner (2009) – that the party of the appointing president and being black significantly influences the fraction of conservative votes. Furthermore, the result for party affiliation is actually stronger compared to what was proposed in the original article, as in our analysis it extends to both, civil and criminal cases.

What explains our different results? We paid particular attention to the code generating the fraction of conservative votes. As multiple reshaping and grouping operations as well as joining of different datasets were necessary in order to obtain this variable, its calculation is not exactly trivial. We imagine that a small mistake in the original code by Landes and Posner (2009), such as an inner instead of an outer join, could change the outcome in the fraction of votes to a significant degree. In turn, its association with the dependent variable may also change. Therefore, we were unable to replicate the exact summary statistics of the dataset Landes and Posner (2009) used as they did not provide their code for replication nor did they sufficiently specify their corrections in the original paper. That, in particular, may affect the differences. In order to extend the dataset, we experimented with different classifying algorithms, where the best one was a passive-aggressive classifier for economic cases, reaching an f1-score of 74.49%. In order to assess the validity of the classification, we compared the regression results obtained by using predicted data to those obtained by using only hand-labeled data. Coefficients found to be significant with the replication as well as with the robustness checks were not replicated with the predicted data, suggesting that 1) the classifier still needs improvement, or 2) researchers should be careful with using predictions as data in downstream empirical analysis. Future research should, therefore, take into account that the distribution of the Songer data in regards to cases per circuit per year does not mirror the distribution of the universe, and as such it may skew the predictions of any classifier. Oversampling is only an imperfect correction for this issue, as is the inclusion of the

circuit or year as a feature. Otherwise, the consistency of results may not be guaranteed.

One aspect that we neglected thus far is that predictions cannot be directly plugged into a regression without correcting for the classification error. Fong and Tyler (2018) proposed one approach to do so. However, Fong and Tyler (2018) describe a case in which one or more independent variables are predicted. In our case, however, we predict the dependent variable. Therefore, we propose to develop a correction approach in order to prevent forward propagation of the prediction error used within a dependent variable which at this point may be of the main reason for failure. Furthermore, the distributions of the enlarged dataset and that one of the original data are significantly distinct. Overall, the classifier was trained on roughly 0.5 percent of data instances when compared to the number of labels that were predicted. As soon as such a considerable dissemblance is present, non-random draws or the lack of stratification is very problematic. Lack of stratification is the case with the original Songer database, i.e. Songer (1993) does not keep the original distribution of cases per circuit as they focused on preserving other aspects such as the presence of all circuits in each year.

Taking the above into account, our results provide a concise groundwork for future research in this area. First, in order to establish a ground truth that goes beyond mere statistical significance and also looks at distributional aspects more than just regression results are needed. Here, we suggest that taken our results multiway error component modeling as well as an extreme bounds analysis should be used on any prior results before trying to take them as a baseline for any extension of the Songer database. Secondly, in regards to machine learning, we show quite clearly that any input which does not include the complete opinion text in some form cannot result in a good overall performance. That is important as it shows that other aspects which are otherwise very useful in the domain of law, such as citations for citation networks, do not contain enough information for this specific task. This holds despite the fact that when using citations as input, the classifier uses many citations to which it assigns the correct ideology label if one were to label them by hand. However, when taken as an aggregation, neither citations nor quotations are distinctive enough. Moreover, while the Songer database features four labels, our results show that the error the classifier makes on the "mixed" label is nearly equally split between "conservative" and "liberal". As the "other" label is negligible in terms of occurrence, we can, therefore, conclude that training a classifier only on the two labels "conservative" and "liberal" does not introduce any systematic. Due to the increase in performance, such a setup should consequently be preferred. Lastly, looking at the regression results, it may be that text alone is not enough. Future research should therefore also think about taking meta-information, such as the circuit court it was heard at, into account. Moreover, looking at the literature of the median judge (e.g. Andrew D. Martin, Quinn, and Epstein, 2004) it may also be important with which other judges a judge sits on a panel. This may be another important aspect, a machine learning classifier may have to take into account.

We hope that our work acts as a baseline on which future work can build on. The obvious next step is to scale back on the interpretability of the model in favor of sophistication: Specifically, we propose a modified doc2vec model in combination with an attention mechanism. Furthermore, future work could stack multiple classification algorithms tailored more closely to the rules of the coding book that the Songer database provides. Another exciting avenue for future work is to compare in-depth the differences, advantages, and disadvantages of various methodological approaches. A particular exciting comparison is a Bayesian framework, as proposed by Andrew D. Martin and Quinn, 2002, compared to machine learning approaches, as suggested by this paper. Apart from methodological extensions, a more content-related one is particularly interesting: Most of the literature is targeted towards high ranking courts, such that the Supreme Court or Circuit Courts. This lack of attention towards lower courts might stem from the fact that the universe of cases to code is vast. Consequently, not even a partially coded dataset, as far as political ideology labels are concerned, is available for lower courts. A classifier trained on Circuit Courts' opinions could predict the label for opinions of lower courts and, by that, help to close this particular gap in the literature.

Chapter 4

Code is Law

How COMPAS Affects the Way the Judiciary Deals with the Risk of Recidivism

Abstract: In the US, many judges receive a prediction of defendants' recidivism risk generated by the COMPAS algorithm. If judges implement the prediction, they delegate a normative decision to proprietary software. Thus far, the debate around COMPAS has focused on (racial, age, and other) discrimination. Using the ProPublica dataset containing defendants' features and the associated COMPAS predictions, we show that the normative concern grounds even deeper. At face value, it predicts which defendants would need to be detained in order to reduce the risk of them committing new crimes before they are brought to justice. Those predictions favor imprisonment over release, hence COMPAS is biased against the defendant. By deciding on such a trade-off, the software provider assumes a role that democratic constitutions reserve for Parliament. Further, we not only show that this bias can be removed, our proposed correction also increases the total model accuracy, and attenuates the anti-black and anti-young bias inherent in the prediction. However, it also slightly increases the risk of a false negative decision. Thus, we show that design decisions regarding the specifications of the algorithm as well as the presentation of its output may be the actual drivers of the aforementioned biases. Based on these insights, we argue that the normative decisions hidden in the design of the algorithm must be made transparent, and that legislators and judges must be enabled to adapt the algorithm to their normative convictions.

4.1 Introduction

Judges have the power to make potentially life-altering decisions affecting any individual. They do not only engage sovereign powers, but they also authoritatively remove uncertainty. The need for such authoritative intervention is patent if the law itself is to react to uncertainty. Hence, there are legal rules that condition interventions on a prediction. However, by definition predictions always include a certain likelihood to be erroneous, as may be illustrated by the the prominent choice between bail and jail. The fact that a person has been apprehended for purportedly committing crime is a predictive, behavioral signal and in general, persons with a criminal history are more likely to commit crimes than those without (Sampson and J. H. Laub, 1992). Yet, the constitutional presumption of innocence forces the law to strike a balance between incapacitation in reaction to the signal, and curtailing the freedom of a person who might not have recidivated while waiting for trial.

For a computer scientist, this is a familiar choice: the incidence of false negative predictions can usually only be reduced when increasing the risk of false positive predictions, and vice versa. This is a normative decision and for the concrete case, life, limb and property of innocent victims are at stake if false positive decisions are minimized. If false negative decisions are minimized, innocent defendants risk loosing their jobs, families, and being put on a criminal career (Hagan and Dinovitzer, 1999; Western, Kling, et al., 2001; Western, Lopoo, et al., 2004).

The legal system cannot avoid making this choice. It notably also makes a choice if it seeks to maximize accuracy, meaning it minimizes the sum of false positive and false negative decisions. The strategy for how that sum is to be minimized, for example by prioritizing one error over the other, is not automatically defined. Hence the policy question cannot be whether this choice is made, but how. It would be simple if this decision could be logically derived from first principles. However, most societies do not feel comfortable with putting a price tag on life, limb, or fear, nor is there an agreed cost for wrongful conviction or suspicion (Brooks and Simpson, 2012). Hence, even if one were to agree on a utilitarian norm there would be disagreement about parameters. Moreover, it can by no means be taken for granted that the well-being of victims and potentially innocent defendants should be traded against each other. It is so, as from a deontological perspective, the freedom of a person from intrusion on her physical well-being should deserve absolute protection, as should the freedom of a person from unjustified sovereign intervention.

As the normatively correct decision cannot be found by deduction, decision-making power becomes critical. In a democracy, the natural institution for this kind of value judgment is the parliament. Possibly, the constitution wants to convey at least some of this authority to the judiciary. As a matter of fact, this happens in the frequent situation of statutory provisions leaving room for interpretation. By contrast, for obvious reasons corporations are no first-order rulemaking bodies: they lack democratic legitimacy. For pragmatic reasons, legal orders make exceptions. Private ordering is, for instance, frequent in the formulation of technical standards. But at the least, such secondary rulemaking

bodies are exposed to scrutiny by institutions with direct democratic legitimacy, like Parliament or regulatory agencies controlled by government.

In 1999, scholars working at the intersection of law and computer science alerted the public to an emergent phenomenon: code is law (Lessig, 1999). Originally, however, the attention was on technical substitutes for traditional private ordering, like the design of a negotiation platform. In this paper we argue, and empirically demonstrate, that normative decisions at the core of constitutionally protected freedoms are now buried in code. This is alarming as these decisions are not at all transparent. To this end, we look at the at the "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) system. At face value, COMPAS provides judges just with computer-generated advice on different aspects of a defendant, such as risk of recidivism or the risk of failing to appear to scheduled court hearings. For our purposes, we focus on the risk of recidivism.
The company, Northpointe, immunizes itself from criticism by insisting that its software only assess the risk of recidivism (general or violent) but does not automatically decide on what is to happen to the defendant. At the surface, it is left to the individual judge what to do with this advice. Yet, unless judges plainly disregard the machine generated prediction, it has the potential to influence their decisions (Grgić-Hlača, Engel, et al., 2019). This potential for influence is normatively highly problematic, as we show that the output of the COMPAS system is influenced by (normative) considerations in the design. We find that COMPAS strongly privileges victims over defendants and this decision may even be in line with the preferences of the majority of the legislator in at least some of the US states. Critically, however, these legislative bodies themselves have never made this decision, the public has never gotten a chance to discuss the choice – it is hidden in the design of the algorithm.

As COMPAS is used in an area of high relevance for individuals, it was met with considerable criticism. In particular, the public debate has focused on hidden discrimination, by race (Angwin et al., 2013; Fass et al., 2008; Flores et al., 2016; Dieterich et al., 2016; Chouldechova, 2017; Agarwal et al., 2020), or by age (Rudin, 2019; Jackson and Mendoza, 2020; Rudin et al., 2020). It has been pointed out that the accuracy of COMPAS predictions is as low as 68% (Grgić-Hlača, Zafar, et al., 2018; Beriain, 2018). Moreover, critics oppose the proprietary nature of the tool (Freeman et al., 2016; Carlson, 2017; Beriain, 2018; Nishi, 2019), with the ensuing potential for conflicts of interest (Freeman et al., 2016) and the lack of transparency (Rudin, 2019). Moreover, COMPAS predictions are no better, in terms of accuracy, false positives (FP) and false negatives (FN), than untrained human laypersons (Dressel and Farid, 2018), at least under comparable circumstances (Jung et al., 2020).

While the concerns raised in the literature are normative, the debate itself has obscured an even deeper normative issue: COMPAS influences judges to the detriment of defendants. In our analysis, we not only show that this bias is pronounced, we also present a technically relatively easy procedure for removing this bias. In line with our earlier call for democratic legitimacy, we do not argue that this corrected version of the algorithm is preferable. However, we want to show that a correction is possible leading to a comparable outcome for which other subgroups profit. Through that we want to

emphasize the importance of discussing the implications for the conflicting normative goals, before legislators allow the use of COMPAS in court. And, if a legislator approves of the bias against defendants, the bias should at the least be made transparent. Hence, that trade-off has to be at the core of a debate about the constitutional limitations which at present lies beyond the scope of this paper.

4.2 Method

There already is a mature literature using the so-called ProPublica COMAPAS dataset (Angwin et al., 2013) for their research (e.g., Angwin et al., 2013; Fass et al., 2008; Flores et al., 2016; Dieterich et al., 2016; Chouldechova, 2017; Agarwal et al., 2020). ProPublica is an NGO interested in promoting fairness and transparency when the state interacts with individuals. For that reason, they made use of the freedom-of-information act to compile a static, dataset of COMPAS scores and the characteristics of individuals to whom these scores correspond. However, the dataset provided by ProPublica is closed-source as well, meaning that one cannot infer how they arrive at their variables from the raw data. Moreover, ProPublica also do not provide the code necessary for that transformation. Consequently, we opted to work with their raw data and make use of the code by Rudin et al. (2020) to transform the raw data to the COMPAS dataset. We outline the the dataset used as well as the design of the correction model in the following.

Data

Our data consists of 5,759 observations from defendants who have been tried in a single county, namely Boward County, Florida. Relying on freedom of information legislation, this data has been collected by ProPublica (Angwin et al., 2013). Further features which are used by the COMPAS algorithm but not available in the original ProPublica dataset, have been added by (Rudin, 2019). Consequently, we know for each defendant priors such as whether she has been incarcerated or released on bail; whether she has been charged for any other crime during two years after release,[1] as well as 32 more directly observable characteristics, mostly demographic and concerning the defendant's criminal history.[2] We do not have access to the remaining features COMPAS uses which are answers to questionnaires on the privatethe screening process (Northpointe, 2015). In terms of distribution, little more than a third (2079) of the defendants in the sample recidivated, in terms of race 2939 individuals are black, 1934 white, and the remaining 886 are of a different race.

Judges are informed about the recidivism risk of a defendant with the help of decile scores calculated by the COMPAS software. These scores result from partitioning the data

[1] Except if the new charge was a traffic ticket or a minor municipal ordinance violation, failure to appear in court, or a later charge with a crime that had occurred before the COMPAS screening (Larson, Jeff and Mattu, Surya and Kirchner, Lauren and Angwin, Julia, n.d.). For defendants who have been in jail, the time until recidivism is measured from the day of release from prison onward. This imbalance is inherent in the data.

[2] The information about the features is summarized in Table D.1, Table D.2, Table D.3, and Table D.4 in the Appendix.

into 10 equally sized bins, conditional on the fraction of the population that COMPAS has used for normalization. By the definition used to construct the norm groups, COMPAS predictions may be stratified by a combination of the following: gender, prison, jail, parole, and probation (Northpointe, 2015, p.11). However, we do not know how the decile scores in our dataset were normed, i.e., what norm groups were applied. As may be seen in Figure 4.1, the underlying raw scores are not exclusive to one decile but rather overlapping at the boundaries – sometimes even the boundaries of two consecutive deciles (as it is the case for decile 3,4, and 5). As such, it is fair to assume that the underlying norm groups are indeed different ones for different subgroups. Moreover, the interpretation of these deciles is then very difficult when two individuals whose score was normed against different groups should be compared. To prevent this additional step from biasing results, we work with raw scores. In order to gain COMPAS decile scores, we apply the COMPAS uniform bracketing to the raw scores in the dataset. In other words, the norm group is the complete training dataset. The deciles themselves are constructed as outlined in

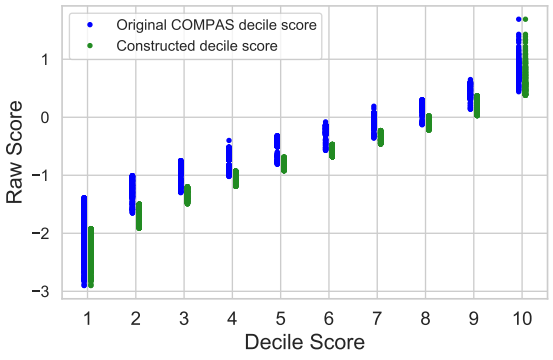


Figure 4.1: Original mapping of raw scores onto decile scores vs. constructed mapping. The constructed mapping is for one norm group and the binning used is the uniform binning.

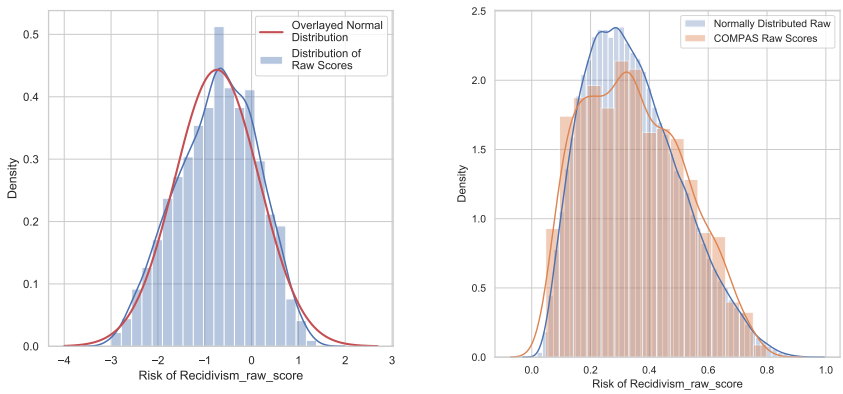
Decile Score	Raw Score Upper Bound	Risk Upper Bound
1	−1.92	12.8
2	−1.5	18.2
3	−1.2	23.1
4	−0.93	28.3
5	−0.69	33.4
6	−0.47	38.5
7	−0.23	44.3
8	0.02	50.5
9	0.37	59.1
10	1.69	84.4

Table 4.1: Raw score and risk cutoffs per decile for data preprocessing. Risk is calculated as the sigmoid transformation of the raw score. The binning used is the uniform binning

the practitioner's guide (Northpointe, 2015), i.e., as 10 brackets, each one capturing one tenth of all individuals. For the final processing, we take the uppermost raw score of each decile (see Table 1) as the threshold for the bin and then resort all individuals in such way that each individual is in the correct bracket. While that means that the bins are not perfectly uniform in distribution, the effect is negligible in terms of impact on the data.

As the left hand panel of Figure 4.2 shows, the raw scores are approximately normally distributed, with a mean and mode around -1. However, the raw scores are not easy to interpret. Moreover, the decile scores are also not easy to interpret in a way that is useful for the direct inference of the predicted likelihood of recidivism. The only information we are able to find is a note in the practitioner's guide that the COMPAS

risk scales are a "method of estimating the likelihood of re-offending" (Northpointe, 2015, p. 29). Thus, we apply the common sigmoid[3] transformation on the risk scores (Niculescu-Mizil and Caruana, 2005); this enables us to interpret the scores as a likelihood. In the right-hand panel of Figure 4.2, we superimpose such a sigmoid transformation. These scores, indicated in orange, can be interpreted as the predicted probability of recidivism. Hence, for the average defendant in the dataset the predicted (raw) recidivism risk is substantially below 50 %. These are also the estimated risk probabilities used in Figure 4.4. While this assumption is only an approximation, it is notable that our results hold even under the unfavourable assumption that the COMPAS raw scores are very closely related to the real risk of recidivism.



(a) Raw Scores (b) After Sigmoid Transformation

Figure 4.2: Distribution of recidivism scores in the dataset.

The figure shows the distribution of the raw scores within the dataset, as well as how normally distributed raw scores with the same bounds (location = mean of raw scores, scale = 0.9) would compare. Figure b shows the distribution after applying a sigmoid transformation to gain probabilities

When performing our correction, we randomly split the data 75/25 into a remainder set and a test set respectively. The test set therefore holds 1440 samples. Then we split the remainder set again 75/25 into the actual training set (3239 samples) and the validation set (1080 samples). We train the classifier on the training set, select model parameters based on validation set error, and report the results on the test set, which is left untouched until the final analysis.

[3]The sigmoid transformation is equivalent to the logistic transformation commonly used to generate probabilities as regression output. It transforms a range of numbers such that the transformed data lies between 0 and 1.

Model

The models for each threshold, to predict errors in the COMPAS predictions, were specified as four-layer neural networks, each layer consisting of 28 neurons using a ReLU activation function. As input, we normalize the features contained in the dataset between -1 and 1. Additionally, the network receives information p about whether the instance would be considered a positive prediction or negative one in terms of recidivism, dependent on the COMPAS decile score d and the threshold t for which the network is trained. Hence, p is defined as follows:

$$p(d,t) = \begin{cases} 0 & d < t \\ 1 & otherwise \end{cases}$$

Our model M is trained to predict the errors of COMPAS. In order to construct these errors, we use the ground truth information g on whether an individual recidivated within the last two years, with $g = 1$ if that is true and $g = 0$ otherwise. Consequently, the COMPAS errors e are defined as follows:

$$e(p,g) = \begin{cases} 0 & p(d,t) = g \\ 1 & otherwise \end{cases}$$

Finally, our ex-post correction model is specified on the input features x, with x being a feature from Table D.1, Table D.2, Table D.3, and Table D.4 excluding the features pertaining to race. As our target, we try to predict the individuals for which the COMPAS assessment would be erroneous. That means, given a threshold t, we try to predict e:

$$\hat{e} = M(x,p) \tag{4.1}$$

The loss function we use is the mean squared error loss $MSE(e-\hat{e})$. The model optimization is done over 250 epochs with a batch size of 500, using the Adam optimizer with a standard learning rate of 10^{-3}. We optimized the number of epochs as well as the number of the neuron in the hidden layers making use of the validation set. However, results vary only minimally before and after optimization.[4]

4.3 Results

Anti-defendant bias from low accuracy

At its face, COMPAS leaves potentially contentious normative choices to its judicial users. The user manual stresses that it is for the user to define, which recidivism prediction to consider problematic, and to react, for instance by denying release on bail (Northpointe, 2015, p.5). The manual also explains that the scores are relative to the group of the

[4]We started with a network having two layers and as many hidden neurons as we have input variables (30). The final network has four layers and 28 hidden neurons.

population to which the defendant belongs (Northpointe, 2015, p.11). These so-called norm groups differ by gender, and whether defendants in the training data have been in prison or on parole; in jail; or on probation (Northpointe, 2015, p.11). Compared with these norm groups, a defendant with a score 1-4 is considered low risk, 5-7 medium risk, and 8-10 high risk (Northpointe, 2015, p.8).

Yet, COMPAS is actually pronouncedly normative, to the detriment of defendants. For the individuals in our data, we know the ground truth. Thus, comparing it to the COMPAS score, we can identify the complete confusion matrix for a given t. We thus know who has been correctly and who has been incorrectly classified for any chosen threshold of the COMPAS decile scores. The left panel of Figure 4.3 shows, on the testset drawn from the ProPublica data which is comprised of 1440 samples, how strongly this assessment is biased. Effectively, with the score which the judge uses as threshold, she not only decides about the acceptable recidivism risk. She also decides how often the prediction is wrong, to the detriment of society at large, or to the detriment of the defendant. If the judge decides to intervene whenever the score is equal or above 4 (medium risk in COMPAS' classification), 577 (40.1%) defendants are wrongly classified as high risk, while only 73 (5.1%) are wrongly classified as low risk. By contrast, if the judge draws the line at 7 (high risk in COMPAS' classification), the incidence of false positives (239 or 16.6%) and of false negatives (275 or 19.1%) is almost balanced. And if the judge aims to be as accurate as possible, the threshold at 9 would be desired.

Importantly, the anti-defendant bias inherent in choosing a low threshold is not communicated to judges. The user manual exclusively focuses on recidivism risk. The judge is led to believe that risk aversion regarding the community is all that is at stake. Yet, as the left panel of Figure 4.3 shows, at a threshold of 4, there are even more false positives than true positives (TP): more than half of the defendants put into jail have not recidivated in the two years after the assessment. About a third of all the defendants are unnecessarily incarcerated.

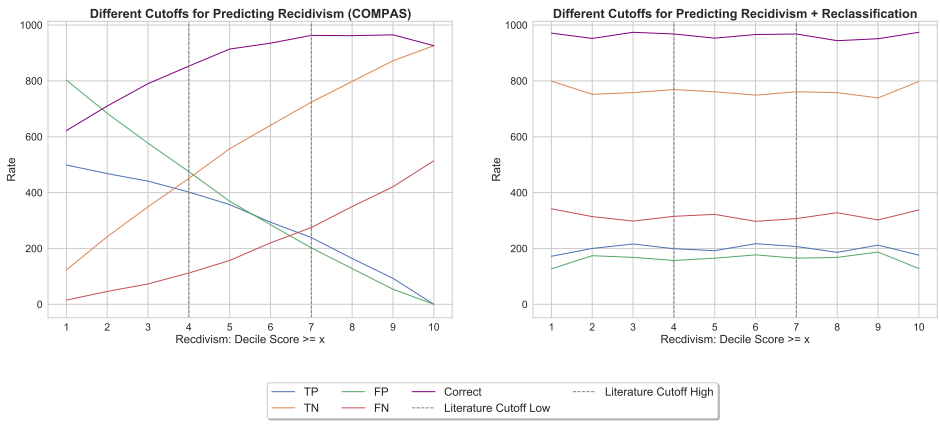


Figure 4.3: Bias against defendants. Left panel: COMPAS, right panel: with ex post correction.

Removing the bias

One may wonder whether the anti-defendant bias is mechanical: does the frequency of false positive choices (defendants unnecessarily incarcerated) not automatically increase if the legislator, or the judge for that matter, is more concerned about innocent victims? Even then, it might still be important to inform judges about the effect. They should be aware that they are exchanging an abuse of sovereign powers against greater safety in the community. However, this trade-off is seemingly inherent in the uncertainty about the recidivism risk of individual defendants.

The right panel of Figure 4.3 shows that this is not quite true. On the x-axis, we depict the chosen decile threshold from which onward a judge considers a defendant for jail. On the y-axis, we show the absolute of number of individuals affected for the different groups, i.e., false positives, true positives, false negatives, and true negatives (TN). We see that the bias can largely be removed using the ex-post correction model specified in subsection 4.2. At the same time, we still leave it to the judiciary to fix the threshold, and hence the acceptable estimated risk of recidivism. In this ex-post corrected version of the COMPAS outcome, all lines are nearly flat. Our correction not only makes predictions more accurate, it also makes the decision for a threshold and the incidence of materially wrong decisions orthogonal. Society may leave it to judges (or the legislature, for that matter) to define the preferred balance between protecting victims and protecting innocent defendants. This choice no longer affects the expected frequency of wrongful judicial decisions. That in itself is, again, a normative decision - we prioritize accuracy over preserving an implicit error-rate, seen as acceptable by the judge. Now, mostly irrespective of the cutoff, more than 2/3 of all decisions are correct. This is undoubtedly a desirable property. Of the 926 defendants who have actually not recidivated, on average 761, i.e. 82.18%, are released, which is desirable. However, of the 514 defendants who have recidivated, on average only about 201 (39.11%) are incarcerated. Considerably more than half of them are released. The judiciary may deem this risk too high. We do therefore not argue that the corrected version is superior. But we show: even when only exploiting the relatively small sample published by ProPublica, the normative trade off between protecting innocent defendants and protecting innocent victims cannot only be made visible, the incidence of wrongful incarcerations can even be minimized. If the judiciary decides not to do so, that in itself is a valid choice. However, that decision, in turn, should be made transparent. Moreover, it should be politically discussed and justified and potentially weighted.

COMPAS does not only hide the anti-defendant bias resulting from the way it handles accuracy computation. A further normative decision is concealed in the way how COMPAS partitions the data, as the output of the COMPAS model is not directly given to the judges. Rather, some post-processing is applied before. As such, in the following we look on that post-processing.

Anti-defendant bias from partitioning the data

The COMPAS model itself does not output the decile scores we have looked at so far and which judges see. Instead, the model outputs raw scores for each defendant. Those scores correspond to a predicted risk of recidivism on an unknown metric scale. Then, the data is split into 10 bins. The average risk score increases from bin to bin, however, the data is not binned by the risk score itself. Rather, in each bin there are equally many data points (Northpointe, 2015). COMPAS explicitly leaves it to the user to decide either for an extended supervision or for an incarceration and only states that each bin can be translated into the percentage of individuals of the comparison group who are more or less dangerous compared to the defendant. That means, for a defendant with a decile score of 4, 60% of people in the norm group are more dangerous. However, COMPAS decile scores could only be translated into predicted probabilities of recidivism if these scores were uniformly distributed across the probability space, i.e. between 0% and 100% predicted risk. The reason obviously is that the distribution across deciles is uniform, thus only if the distribution of individuals across the "risk of recidivism"-range is uniform as well does the risk increase linearly across deciles. That important fact, however, is – to our knowledge – not mentioned at any point in the practitioner's guide (Northpointe, 2015) nor explicitly stated to the judges. Hence, the user does not have any information about the distribution of the risk scores in the norm group, and subsequently, she is unable to asses whether an increase of the decile score by 1 corresponds to an increase in predicted risk by 10%, of more, or of less.

As Figure 4.4 shows, the reality is very far from such an increase of 10% and does not even correspond to the same increase for each decile. As the blue bars show, the actual distribution is heavily skewed to the right: there are many more observations with a low risk of recividism. Actually, if the judiciary adopts the "medium risk" threshold often proposed at COMPAS decile 4, the maximum estimated recidivism risk is not 40 %, but 28.3 %: of the most risk prone defendants at this borderline, only about 1 out of 4 is expected to recidivate during the next two years. COMPAS decile scores are heavily biased, again to the detriment of defendants.[5] In the field guide, Northpointe singels out COMPAS deciles as corresponding to the "x%" most dangerous defendants (Northpointe, 2015). This might be what the legislator aims for. But the legislator should be aware that this definition of the threshold, if falsely interpreted as probability thresholds, creates an additional bias to the detriment of defendants who very likely would not have recidivated when released. This normative decision, too, should not be concealed. For transparency, COMPAS should offer alternatives for the decile binning strategy. After the effect has been made transparent, it is for the legislator to defend its choice.

[5]The bias is even stronger if one were to partition the COMPAS raw scores into 10 bins of equal *width*. One sees this in Figure 4.4: For all predicted probabilities of recidivism below 70%, the red line (raw scores) is above the orange line (probabilities).

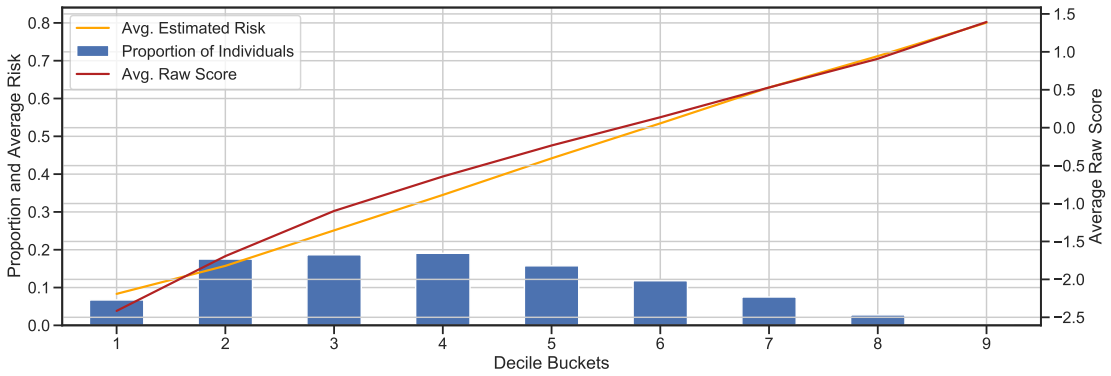


Figure 4.4: Alternative definition of cutoffs.

The figure shows an alternative for partitioning the data: for the decile boundaries calculation, first the predicted COMPAS raw score for recidivism is transformed by a logistic/sigmoid function. Then the 0-1 scale is split into 10 evenly spaced bins and the defendants are sorted into them using the transformed raw score. Blue bars show the frequency of observations, in the respective bin, in the COMPAS data. The orange line stands for the average sigmoid-transformed, risk raw scores of all observations in this bin. The red line stands for the average COMPAS raw score for all observations in this bin.

Racial bias revisited

Triggered by ProPublica's findings, the normative debate has been focused on racial discrimination (Angwin et al., 2013; Fass et al., 2008; Flores et al., 2016; Dieterich et al., 2016; Chouldechova, 2017; Agarwal et al., 2020). Figure 4.5 casts new light on this finding as it show how the rate of the false-positives and false-negatives differs for changes in cutoffs. We show that for the original COMPAS predictions as well as the ex-post corrections when applying our model. It shows that the races are not symmetrically affected by the risk of being unnecessarily incarcerated (left panel) or by the chance to be released on bail without justification (right panel). Irrespective of the cutoff, black defendants are substantially more exposed to false positive rulings, and white defendants are substantially more likely to benefit from false negative rulings. The ex-post correction introduced above is also effective conditional on race (the dashed lines are in parallel). Despite the fact that the correction is tuned towards accuracy and does not directly interact with race, the racial bias is clearly reduced after the correction.

Age bias revisited

In recent years, academic attention has shifted from race to age (Rudin, 2019; Jackson and Mendoza, 2020; Rudin et al., 2020), chiefly because it has been recognized how both variables are correlated. Black defendants have an average age of 29.53, whereas white defendants are 35.15 years old on average. Hence, black defendants are much younger on average.

For lower decile scores, and therefore especially for the thresholds 4 and 7 popular in the literature, compared to COMPAS, the correction model reduces the maximum

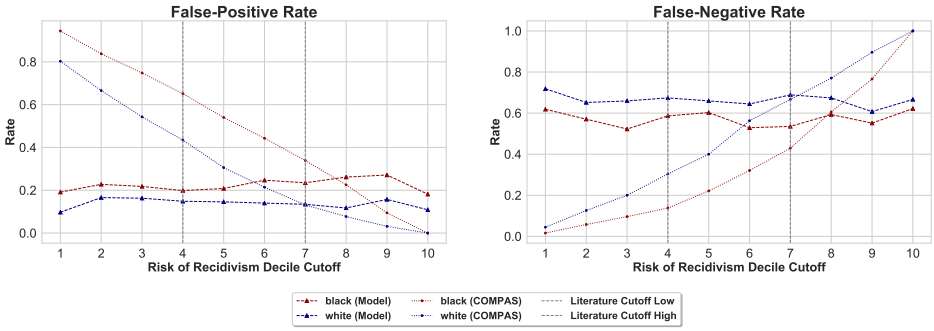


Figure 4.5: Racial bias in false positives vs. false negatives.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel). Dotted lines: results when using COMPAS predictions, conditional on threshold chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Red: black defendants, blue: white defendants. Other races are excluded.

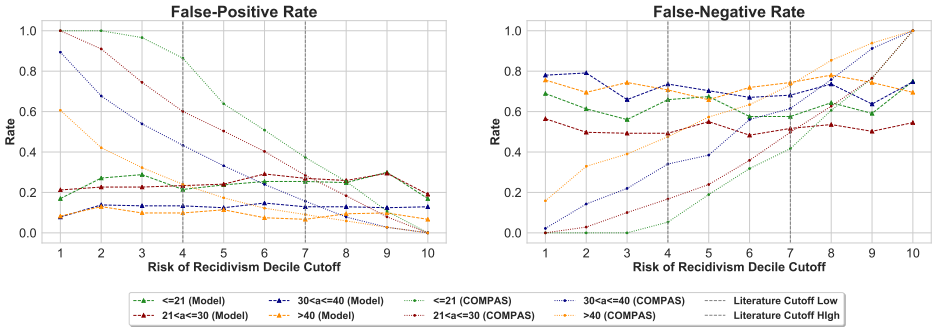


Figure 4.6: Age bias in false positives vs. false negatives.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel). Dotted lines: results when using COMPAS predictions, conditional on threshold chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Green: ≤ 21, red: $(21, 30]$, blue: $(30, 40]$, orange: > 40.

difference of error rates between groups. As shown in Figure 4.6, the strong bias against young defendants actually results from a much higher risk of false positive decisions (the younger the defendant, the higher the dotted line on the left panel). This corresponds to a considerably lower chance for younger defendants of being released on bail without justification (the younger the defendant, the lower the dotted line on the right panel).

For the race variables, the accuracy correction introduced above, partially corrects the bias (dashed lines in Figure 4.5). For the age bias presented in Figure 4.6, the correction model can directly make use of the bias variable. Thus the correctional effect is even stronger. After correction, defendants between 21 and 30 years are disfavored, both in terms of a relatively higher false positive rate, and of a relatively lower false negative rate. Still all age brackets benefit unless the decile score threshold is set at 8. Moreover, the bias against the different sub-groups shown resulting from the choice of the decile score threshold is neutralized (also conditional on age, the dashed lines are flat). As the model does not get information on race as input but age (same as COMPAS), we assume that the race bias generated by the model is a side effect of an explicit age bias. Thus, when the latter is corrected the former also undergoes a partial correction.

4.4 Discussion

Naturally, it is not preferable that human employees add up hundreds of numbers; computers are just better at this task. In the field of medicine, for example, many doctors have embraced evidence based medicine (Sackett et al., 1996); at least for some diagnostic tasks, dedicated software outperforms human experts. It does not seem far-fetched to draw the analogy to judicial decision making, and to call for evidence based adjudication (Manski et al., 2020). The more the law wants the decision to rest on a prediction, the more it seems appealing to muster the ever increasing capacity of algorithms, paired with the growing richness of datasets, to make good predictions. In this paper, we do not argue against computerized decision aids in the court room. But the law does not only care about performance. The fact that, on average, in a given domain, one decision maker (the machine) is better than another (the human judge) does not automatically imply that this decision maker should decide. Ultimately, the law does not care about population effects; it cares about individual cases. Now in many contexts, the law must live with imperfection. Facts that are relevant for the decision of the case remain unclear or are contested. The risk of materially wrong decisions is insurmountable.
Traditionally, the law contains this risk procedurally. The most important safeguard is the personality of the judge. She is obliged to weigh the pros and cons as best she can, and to be responsible in person for the outcome. To the extent that concerns can be generalized, the legislator takes a stand, based on open political debate. The rule of law and democratic legitimacy make the residual imperfection tolerable. In this paper we show that these safeguards risk being blunted by algorithmic design. The popular COMPAS software does not only perform poorly at the very task for which it is has been designed; this has been pointed out before. The software also implicitly assumes the role of a sovereign that, in a democratic country, should be with the courts and the legislator.

It pronouncedly privileges potential victims over potentially innocent defendants. More importantly: this normative decision is concealed. The judge is led to believe that she just decides about the acceptable recidivism risk, while she actually also decides about the risk of unnecessarily incarcerating a defendant, and about racial and age bias.
Noticing the hidden normative dimension of the algorithm is not an argument against algorithmic decision aids in the court room. However, it is only due to ProPublica's efforts that COMPAS predictions can be compared to ground truth. And with this analysis, we are the first to show the hidden bias inherent in the design of the COMPAS software. Legislators and (constitutional) courts should only clear the use of prediction software in the courtroom if such normative decisions have been made transparent. The algorithm should be tested against ground truth. Following our approach, corrections should be developed. It would then be for democratically legitimate authorities to make the inevitable normative decisions. Our study had to live with the available data. The ProPublica dataset is only a sample, for one jurisdiction. We also did not have access to the complete feature set that COMPAS uses for prediction. Hence our correction algorithm cannot exploit the full richness of the data, but as it performs very well regardless, this does not seem to be an important limitation. As we have shown, code is law. If code is used in the court room, proper safeguards for the rule of law and democracy must be developed.

CHAPTER 5

Systematic Errors and the Stability of Feature-Relevance

An Assessment for the Social Scientists

Abstract: Machine learning sees ever-wider use, especially in the social sciences as well as law. Here, Natural Language Processing (NLP) finds many use-cases as it enables the processing of large-scale online text data. However, often the focus only lies on a model's performance and not on its transparency. That lack of transparency is especially troubling as the users of those models may often not be its designers. As such, they have no way to assess whether there are systematic errors or whether the performance hinges on unknown, possibly unstable factors. Here, we offer an in-depth analysis of the effect small variations in the input have on systematic errors and feature-stability. Our aim is to enable social scientists and practitioners using the technology to assess whether they may invite either normative or unwanted systematic errors into their results when using current technologies.

JEL Codes: C71, H41

5.1 Introduction

In traditional studies within the social sciences, characteristics such as age or gender are key traits as they have proven to be central to understanding and modeling human behavior (e.g., Lahey et al., 2000; Gneezy and Rustichini, 2004; Charness and Gneezy, 2012; Booth and Nolen, 2012; Sutter and Glätzle-Rützler, 2014; Bian et al., 2017). While traditionally studies within the field focused on lab experiments as well as questionnaires, large-scale datasets have, until recently, been of limited availability. Especially when it comes to text data, the analysis often proved resource-intensive and the results are difficult to assess. However, in general, that data is a wealth of information, even more so since the amount of text data generated by individuals increased massively with the advent of services such as Facebook or Twitter (InternetLiveStats, 2019). With such large amounts of data, new methods in machine learning and natural language processing (NLP) gained great popularity, especially within the social sciences. In order to mine that treasure trove, researchers increasingly turn to the application of NLP to answer prevailing questions in the social sciences (e.g., Bail, 2016; Pavlick et al., 2016; Costa-jussà, 2019; Burley et al., 2020). The technology as such is, depending on the algorithms, comparatively simple to use in terms of know-how when one considers the textual features for identification as well as the difficulty of setting up and training the respective algorithms (Narayanan et al., 2012). By reducing the entrance cost in such a way, this technology certainly has great potential. Here, the section of authorship analysis is becoming an area of special relevance to the social scientists. The reason is simply that, while a large amount of text data is available, characteristics of the individual, such as age and gender, but also their identity, often are not. The main goal of that particular area is therefore to profile characteristics such as age and gender, but also political orientation or even the author's identity from texts written by the individual. As such, while studies within the social sciences, in making use of that new type of data, continue to include these characteristics due to their proven relevance in past research (e.g. Bail, 2016; Colleoni et al., 2014), authorship analysis is used to compensate for the lack of ground truth. That is often done by using a layered approach, first inferring the missing characteristics with a trained classifier, and then in turn using that as input to their research approach (Barberá and Rivero, 2015; Huang et al., 2020). Moreover, these characteristics are not only important as an input for further research, but also as identifying information. In the adjacent field where NLP and behavioral sciences intersect, the law community is making use of such analysis for the targeting and researching of incriminating online behavior, e.g., hate speech (Djuric et al., 2015; Z. Laub, 2019; Zufall et al., 2020). The practical application of authorship analysis also becomes relevant when pursuing offenses. Often, online users do not use their real names, and finding out their identities becomes difficult either because it is not available or because the companies having access to the information are not willing to share it (Stuttgarter Nachrichten, 2017). Consequently, often during such an investigation, forensic authorship analysis is employed to gain additional information on the potential offender. Due to the high volume of data, these processes become increasingly automated and as such the research

field of automated forensic authorship analysis is well established (Rocha et al., 2017).

However, while state-of-the-art methods regularly manage to achieve a high accuracy for authorship analysis (Rocha et al., 2017), some call the field's scientific character into question (Chaski, 2001). This is due to the fact that current research of automated authorship analysis mostly focuses on correct predictions, using a wide set of features which often varies between different papers (Rocha et al., 2017). The focus seems to lie on achieving the best results, showing the viability of automated authorship analysis often to the detriment of rigorous explainability and transparency (Chaski, 2012). This not only generally concerns research employing such models but, when used by law enforcement, it is also directly related to the admissibility of findings from automated authorship analysis in the courts, as transparency and rigor might not satisfy the demands before the law (Chaski, 2012). As such, this topic is very much of normative interest. Explainability and transparency are the central aspects when a decision made by some model affects individuals. Within the field of law, as outlined above, that may affect the individual directly as she comes under suspicion when identified by the machine. In social sciences, they may be more indirectly affected. One way would be when predicted labels on a datasset are used for further analysis. The result from such an analysis may inform policy decisions. However, the error will propagate. As the underlying labels assigned by the machine were systematically faulty, the policy design will be as well. On the one hand, it is therefore paramount that, when exposed to such machine-informed decisions and label assignments, there are no systematic patterns of errors. On the other hand, the features driving the prediction result should be more than just expressions of spurious correlations, which are present in one dataset and absent in the next. Consequently, the technology is in need of further assessment before being used even more widely than is already the case.

Besides understanding the algorithm, such an explainability would require two things: One concerns the topic independence, which means that the features predictive of an author should not depend on the content of the text (Narayanan et al., 2012). The second aspect is rarely mentioned. Most models are trained at one point in time, on one particular text corpus related to one domain. Using the trained model at a later point in time, however, assumes that in the meantime there was no shift in the underlying features used or shifts in the set used by a particular author. If one looks at age, for example, it may be that older people are currently using more grammatically correct language in the online environment compared to younger people (Flekova et al., 2016). Naturally, that is merely a snapshot of the current environment. It does not imply that the next generation features the same pattern, and consequently, any model trained on the old pattern might inadvertently misclassify when confronted with the new pattern. For that reason, it is necessary to assess the stability of individual features used by the classifier, when the underlying data and thus the patterns change slightly. There are only very few forays seeking to address such problems, for example as Azarbonyad et al. (2015) do with their temporal weighting of features. As stylometry is rooted in the humanities and thus the social sciences (Neal et al., 2017), it is surprising that not more efforts have

been made so far to see whether some author characteristics result in some stable topic- and domain-independent features and feature relevance.

Therefore, the central aspect of our approach is helping to answer that question about the stability. We assess the stability in terms of predictions and in terms of feature relevance. We think that such a stability analysis would aid immensely in assessing the rigor of predictions, therefore making them safer to use in the legal context. Moreover, it would also help us to better establish the boundaries of transferability and stability of models and their predictions, which is needed when predictions of such models are used as input for further research. This contribution is therefore interdisciplinary in nature, as it tries to address an issue affecting multiple fields. While the lack of contributions has already been pointed out Rocha et al. (2017), only recently, have there been any notable forays. In general, there has been an effort to make model predictions more explainable (Ribeiro et al., 2016b; Samek et al., 2019). A systematic approach , however, looking at changes when features are systematically varied, is still limited. The study by Koppel et al. (2011) looks at authorship analysis "in the wild" and systemically varies the number of authors as well as the number of features to assess and quantify gains and losses in performance. However, the authors do not focus on feature types and do not extend their study towards analyzing the changes in within the model. Recently, Boenninghoff et al. (2019) showed a method to make a complex model based on a neuronal net explainable. Their approach is limited to their specific model and does not analyze either what the decisive features correspond to, i.e., how much context they encode. In that vein, Sanchez-Perez et al. (2017) is closer to our approach. They also seek to limit topic dependency and focus on feature types. However, their goal is to find a good subset of n-grams for their feature set with high predictive power. The paper closest to ours is the one by Sage et al. (2020). Their analysis is focused on different feature types and the influence of varying n-gram lengths. They systematically vary both in order to find the impact on performance. However, they do not extend their analysis to different input sets and also focus on longer news articles instead of the more common data of microblog texts. Moreover, we also extend that analysis into the domain of stability, assessing whether there are shifts in feature relevance.

5.2 Experimental Design and Data

In order to conduct our stability analysis, we conduct an experiment as used in the field of machine learning by introducing controlled variations to an underlying, given dataset. To that end, we use a fixed setup of machine learning (ML) models and test their internal stability, when they are exposed to these controlled variations.

Synopsis. The dataset used for the experiment is the PAN @CLEF 2019 Celebrity Profiling (PAN2019) dataset.[1] As our goal is to assess the performance and the stability

[1]The dataset may be downloaded from the website of the PAN challenge: PAN Challenge 2019.

of relevance in regards to single features, we try to directly reduce variation present within the authors as much as possible. Therefore, we focus only on authors from one category, namely those dubbed "creator". As suggested in the guidelines of the original PAN challenge, we change the age from a numerical variable to their categorical one consisting of five age brackets. As a prediction target, we select "age" and "gender", two commonly used characteristics in the social sciences. Moreover, in order to exclude further any variation introduced by an imbalanced dataset, we undersample the data in such a way that the genders as well as the age groups are balanced. For the comparison in the author dimension, we create four subsets consisting of 50, 150, 500, and 1000 authors, respectively. The upper limit of 1000 authors reflects the maximum number of authors for whom it is still possible to balance the dataset. Furthermore, we repeat the experiment three times for different minimum lenghts per training instance, as the text lengths were shown to impact classifier performance; in doing this, we hold the model setup constant (Custódio and Paraboni, 2021). The minimum lengths are 100, 250, and 500 characters, respectively. In order to achieve these minimum lengths, tweets from the same author were concatenated. As feature types we use the following ones, sorted in ascending order in terms of encoded context information: DIST, CHAR, ASIS, POS, TAG, DEP, LEMMA, WORD, NUM. For all types, we apply the n-gram ranges found to be useful by prior research (Custódio and Paraboni, 2021). The evaluation is conducted using the 500-score as the performance measure. To assess the stability of features, we use Spearman's Rho for rank order correlation. All evaluations were done on a separate hold-out dataset, the test set. That data was not used during training at any point. In the following paragraphs, we outline our design choices in detail.

Data

While the literature for authorship analysis is abundant and only increased during recent years, there are no easily identified commonly used datasets across a wide range of studies. Comparison between studies is therefore difficult. This is well illustrated by Neal et al. (2017) who list 13 datasets used more than once and a multitude of others used less frequently. However, for Twitter, they list only two. This means that most studies on authorship analysis additionally suffer from at least one of two limitations. Either they focus on authorship analysis while using traditional, longer texts, such as articles or blog posts, or they make use of custom datasets (Neal et al., 2017). The latter are sometimes described as a great challenge in the field of authorship analysis, making replication as all well verification of results difficult (Halvani et al., 2016). The former implies that past studies not focusing on online short text messages are looking at a fundamentally different research problem compared to Twitter texts. At the same time, most automated authorship analysis research acknowledges the fact that use cases for these tools will consist of attributing micro-blog texts to an author (see, for example, Narayanan et al., 2012; Rocha et al., 2017; Spitters et al., 2016). Consequently, the dataset used is from that platform, as its prevalence makes it especially relevant. Moreover, it reflects the text data, in style and characteristics commonly found for chat messages. Especially in terms of length, it is also similar to text data generated during studies and experiments

within the social sciences.

Table 5.1: Statistics of the Dataset

		Avg instance length		Avg tweet length		Avg no. tweets per instance	
	Target	age	gender	age	gender	age	gender
No. of Characters	No. of Authors	No. chars	No. chars	No. chars	No. chars		
100	50	160.30	162.94	109.58	111.92	1.46	1.46
	150	160.52	162.37	107.38	112.94	1.49	1.44
	500	160.78	161.63	109.21	111.86	1.47	1.44
	1000	160.07	160.99	109.28	111.84	1.46	1.44
250	50	313.16	315.73	109.05	112.02	2.87	2.82
	150	313.25	315.58	107.43	112.84	2.92	2.80
	500	313.26	314.66	109.09	111.67	2.87	2.82
	1000	313.30	314.34	109.18	111.84	2.87	2.81
500	50	565.60	568.76	109.12	111.87	5.18	5.08
	150	565.89	568.71	107.48	112.84	5.27	5.04
	500	566.13	567.54	109.15	111.70	5.19	5.08
	1000	566.10	567.38	109.17	111.85	5.19	5.07

Most studies focusing on short-text online media such as Twitter use different datasets. This is due to the fact that the user agreement for the API of this particular platform does not generally give permission to publish a scraped dataset online (Theophilo et al., 2019). At this point in time, we know of three public datasets: Twisty (Verhoeven et al., 2016), ISOT (Brocardo et al., 2015), and PAN (Stamatatos et al., 2015), a yearly challenge tackling different aspects of authorship analysis. The Twisty dataset includes a multitude of languages, making it unusable for this task as it was shown that language has major impact on the results (Halvani et al., 2016). Another problem mentioned before concerns the high number of troll profiles, as well as potential alias accounts (Varol et al., 2017) in an arbitrarily captured dataset. This is sometimes referred to as the ground truth problem (Narayanan et al., 2012). As the research question is focused on characteristics of individual people, this is particularly problematic. For this reason, the ISOT dataset, too, is unusable, as neither the issue of troll profiles nor the problem of double accounts for a single user can be addressed.

To overcome this, a special version of the PAN dataset focusing on profiling celebrities (PAN, 2019) is used. For this dataset it can at least be established that the accounts relate to a real, individual human. Naturally, there may be new limitations, e.g., it may not be guaranteed that celebrities always write their own posts. However, Twitter is more and more considered to be a medium offering the possibility of interacting directly with followers by circumventing the filter, interpretation, and comments of traditional media (thus enabling ”authenticity”) (J.-H. Schmidt, 2014). Consequently, the problem of other people messaging instead of the celebrities themselves is considered minor by the

authors of the dataset when compared to the problem of having unknown fake profiles.

Feature Engineering

In order to use text input for machine learning models, the text has to be transformed into a numerical representation. The chosen representation we call feature type here. Within one feature type, there may be many features. For an example of two words, each may be mapped to a number, so there would be two features.
For automated authorship analysis, one may in principle choose from or combine a wide range of possible features for prediction. The natural approach would be to use word-based features. However, this comes with the limitation that rather than finding features predictive of a certain gender or age, it is more likely that the topic is a latent variable driving the result. As our goal is to control most of the information from outside the feature itself, e.g., topic or other context, the selection has to be more nuanced. This brings us to character-based features. Character n-grams, are based on concatenating characters; in the form of 1-grams they equal uni-grams, i.e., single characters. They are maybe one of the most commonly used feature sets within the literature (Rocha et al., 2017). Spitters et al. (2016) find in their exhaustive review of the literature that most studies employ them in one form or another. This is due to the fact that such character n-grams were shown in multiple studies to perform robustly (Kešelj et al., 2003; Stamatatos, 2009; Peng et al., 2003). Some authors like Forstall and Scheirer (2010) link this performance to the fact that n-grams are very closely related to pronunciation. Moreover, due to the fact that the n-gram length may be reduced, many outside influences which introduce context in terms of topics, text type, and even language as a whole may be removed. For example, the cross-domain analysis by Stamatatos (2013) shows that, compared to traditional word-based features, character n-grams outperform them in terms of cross-domain stability. N-grams were also shown to capture many different features such as punctuation or spelling mistakes. Regarding the length of n-grams, it must be noted that for English, those with a length of three and above are shown to capture content again partially, thus becoming topic-dependent (Narayanan et al., 2012; Spitters et al., 2016). Therefore, not only is the type of feature important when controlling for the relevance of topic and content but also the n-grams themselves are crucial. As such, we have a layered approach, controlling for the feature type, while also varying the n-grams employed within one feature type. In the following, we construct a hierarchy ranging from the type of features mostly removed from content to the ones which partially capture content. In between, we can place those features which are still related to style and structure, but which necessitate a certain amount of text. In terms of the actual feature types as well as range of the n-grams, this study mainly follows Custódio and Paraboni (2021) with the numerical features taken from Huang et al. (2020). It gives us the following types as input in ascending order, when compared on their context-dependency.

Text Distortion. Symbols within text are usually disregarded for the standard approaches of text-based models. However, past research has shown that these features serve as valuable information in the context of authorship analysis (Stamatatos, 2017). For this feature type, all a-z characters are mapped to "*", which only leaves punctuation and other markers. We call this type of feature DIST and apply n-grams $\in [2, 5]$

Character. Character n-grams were shown to capture many idiosyncrasies present in text, while yielding a stable performance (Stamatatos, 2013). Moreover, the amount of context present in the n-grams can easily be adjusted by their range (Rocha et al., 2017). Thus, we include them in the range of [2,5], referring to them as CHAR.

Unprocessed Text. In essence, this feature type is a combination of text distortion and character n-grams. As input, the unprocessed text, including all special characters and punctuation, is transformed into n-grams. The n-gram range is also [2,5]. This feature type we refer to as ASIS.

Part-of-Speech. Part-of-Speech tags capture linguistic style patters and general information such as grammatical classes, e.g. "noun" or "verb". We tag the text by employing the SpaCy[2] tagger. This feature type is called POS and the n-gram range is [1,3].

Language-specific morphological features. Within a language, one is also able to find more fine-grained features related to morphology. Such features concern, for example, the gender of a word, tenses and others. To extract these, the tags generated by SpaCy on its TAG level are used. Following this, we call this feature type TAG and employ the n-gram range [1,3].

Syntactic Dependencies. This type of feature captures structural information, e.g., the use of the passive over the active voice. The dependencies are generated using SpaCy's dependency parser. We refer to it as DEP and the n-gram range is [1,3] as well.

Lemma. Lemmas are essentially word-like features, although the words are reduced to a common, lowercase form. For example, "I'm" would be converted to "i" and "am", while "played" and "playing" would both be mapped onto "play". In such a way, words are captured but not their transformations. We call this feature LEMMA and include it with [1,2]-grams.

Words. This feature type is created by forming the n-grams directly from the words without any preprocessing besides lowercasing, and removing all characters that are not within the A-Z range. Again, we employ [1,2]-grams, the feature type is called WORD.

[2]https://spacy.io

Numerical Features. Huang et al. (2020) additionally suggest numerical features describing the content and the form of the tweet. The feature type NUM is thus comprised of the following attributes: average tweet length, number of URLs, number of dates and times, number of emoticons, number of emojis, as well as polarity and subjectivity.

Preprocessing, Models, and Targets. For the preprocessing, we apply the specific ones outlined above to each feature type. In general, emojis and emoticons were always counted as one singular feature and marked by an $< EMOJI >$ or $< EMOTICON >$ token in the beginning and end. Furthermore, we replaced the unicode string by the textual description using the package demoji.[3] For all text-based features, we apply count vectorization and tf-idf scaling. The individual features were kept when they appeared in more than 1% of the samples. For the feature type NUM, we apply scaling and centering. For the model, there is the option of linear and non-linear models. While neuronal nets and transfer-learning models gain huge popularity, for our case of authorship analysis, it turns out that simple linear models regularly outperform the more complex ones (Rocha et al., 2017; Custódio and Paraboni, 2021). Moreover, as we also address the social sciences as well as the law community, interpretability and transparency are key aspects. Thus, we focus here on well-researched models which also enable a mathematically global interpretation, as well as attribution of outcome to individual features of the input. While there is a wide range of models employed, the most common ones are a SVM, a logistic classifier, and Naive Bayes Classifiers (Rocha et al., 2017). We test all three of them and use the overall best-performing one for the evaluation. For the SVM, the analysis is limited to a linear kernel as only this type enables us to interpret the weight matrix directly in terms of feature relevance. As target, we selected two author characteristics of high relevance for the social sciences, namely age and gender.

Experimental Setup

In order to assess how different combinations of feature types impact the outcome, we use three different approaches to feed them into a classifier.

1. Baseline: Here, the model gets only one feature type (although with varying n-gram ranges). Thus, it enables us to compare the performance of individual feature types against one another.

2. Cumulated: For this approach, we feed the classifier combinations of feature types such that we combine them in an ascending order in terms of context-content.

3. Stacked: Here, too, we use different feature types as input. However, we first make predictions using individual feature types (as in the baseline setup) and then apply a second classifier, a logistic one, on top, using the predictions as input to predict the target again. That, in essence, is an ensemble approach (Dietterich, 2000) and a variation of the successful DynAA model by Custódio and Paraboni (2021).

[3] `https://pypi.org/project/demoji/`

To assess the stability in performance as well as relevance, we compare the different feature types we introduce a small, controlled variation on the input data. In order to simulate (possible) shifting variations in the patterns of feature use, we vary the number of authors within the dataset. To that end, we construct four subsets from our dataset. Each subset is comprised of a different number of authors (50, 150, 500, 1000). Furthermore, the sets are constructed in such a way that all authors present in the smaller set are also present in all the larger ones. That means the 50 authors from the smallest set are part of all three larger sets as well. We chose that approach in order to increase the number, and thus the potential variation, while at the same time keeping prior information. Moreover, the authors are balanced in gender as well as in age. That is necessary so that the model has no advantage by focusing on one class to the detriment of others. That gives us a cleaner result when analyzing the impact of the individual feature types as well as n-grams.

We conduct the whole experiment three times, varying the input length of the individual text instances each time. As previous research has shown that text length greatly influences the outcome (Custódio and Paraboni, 2021), we construct input instances of different minimum lengths. We do so by concatenating different tweets by the same author together until the minimum length is reached. No n-grams are constructed in such a way that they would contain information from two different tweets. Naturally, when we increase the minimum length, the number of individual training instances declines, as more tweets are needed to form one training instance. As minimum lengths we use 150, 250, and 500 characters. The summary statistics for the dataset may be found in Table 5.1.

In order to have no spillover of information between evaluation and training, we split the dataset into two subsets, training and testing. All training was done on the training dataset, while all evaluations shown here are done on the hold-out test set. For the stacked approach, we split the test again, this time into a validation and test set. Here, the first layer of the model is trained with the training set, which is the same for all classifiers. The second layer is then trained on the validation set. Finally the evaluations are conducted on the hold-out test set. Due to this setup, all classifiers are trained on exactly the same first-layer input in order to increase comparability.

Evaluation Measures

We seek to answer the question of stability in predictions as well as stability on the level of features. For the former, we test the predictive power of the classifier for both targets on different inputs and sets varying in the number of authors. Our analysis compares different feature types and their performance against each other. Their difference lies in what they capture, especially in terms of the amount of context. We also analyze by how much their inclusion improves the model's performance. As the evaluation metric of choice, we use the macro F1-score. The score is an equally weighted mean of precision[4]

[4] $TruePositives * (PredictedPositives)^{-1}$.

and recall[5]. Moreover, on a balanced dataset, the score is nearly equal to the accuracy. The score is bounded between 0 and 1, with 1 being the optimum.
Furthermore, we analyze the performance on the author level. That helps us to test whether the models make systematic errors for specific authors. Such is indeed of importance, as it tells us something about how patterns, found to be predictive for specific target categories, may systematically disadvantage some individuals compared to others. For that part of the analysis, we look at author level accuracy as well as the stability in classifications patterns. For the latter, we evaluate confusion matrices. The results for that analysis may be found in Section 5.3.

The second, central aspect in this study is that of stability on the feature level. The question we try to answer here is by which degree stays the relevance assigned to single features constant, when the classifier input-set used for training is slightly changed. The change introduced here is the increase in the number of authors. What we want to assess is by how much the relevance of individual features shifts when such a change occurs. We developed the following approach: First, we extract the weight matrix of the two relevant models. When all input features are scaled to the same range as well as centered, the matrix contains the information about the relative relevance of each feature when predicting the outcome. As we are using linear models, these weights are global, i.e., the relevance assigned to a feature is the same no matter which individual instance is assessed.

To assess the potential shift in relevance, we rank the individual features in terms of weights assigned. In a second step, we then calculate Spearman's ρ in order to assess by how much the relevance placed on individual input features shifts when introducing a small variation in the underlying data. The reason why this works is because the ranking of a feature directly reflects the weight the classifier places on it. Within our linear models, this is a direct mapping from its relevance to the prediction result. Thus, when this coefficient is 1, the distribution of relevance across features is completely identical; if it were -1, it would be completely inverse. Consequently, we say the stability is high for values going towards 1, while values close to 0 imply that there is no recognizable relationship, and thus very high instability. Moreover, we selected an ordinal measure, as it allows for more latitude. While the absolute values of the coefficients might change (and indeed have to when more features are included), their ordering may still be constant. Consequently, their relevance when compared to each other can still be stable. That means that any stability found here is to be considered the upper bound.

However, when increasing the number of authors, the underlying feature set might also increase and thus the two matrices do not have the same dimensionality anymore. To tackle that problem, we follow two approaches: The first is to expand the smaller matrix by adding columns of ∞ for the missing features. This ensures, that these will always be assigned the highest possible rank in the smaller matrix (the rankings are sorted in ascending order during comparison). We call this the extended Spearman

[5] $TruePositives * Positives^{-1}$.

correlation and it enables us to assess the absolute feature relevance ranking. The second option is to assess only the features present in both matrices, and therefore to reduce the dimensionality of the larger one. This ensures that we assess relative relevance, but ignore that additional features might have great influence on the outcome. That we refer to as the reduced Spearman correlation. The result for the analysis of the feature relevance and its stability may be found in Section 5.4.

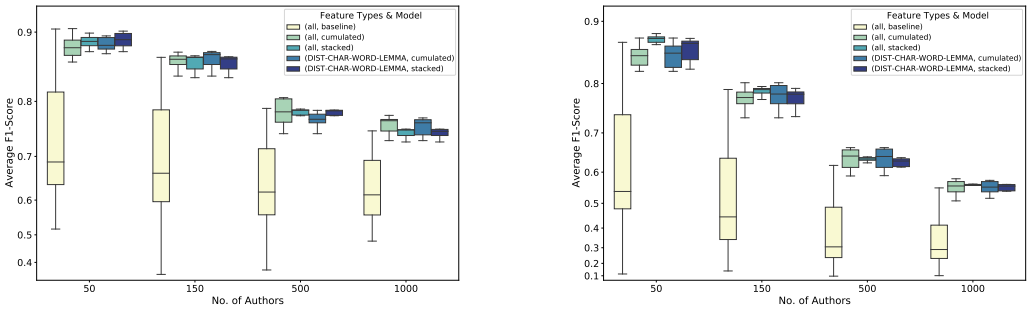
5.3 Stability of Predictions

The results were computed on HPC system, making use of 72 cores with 256GB of RAM.[6] Of our classifiers, the SVM outperformed the logistic classifier as well as the Naive Bayes Classifier as a first-layer model. Thus, for all results a SVM was used as the first-layer model. For the stacked model, a logistic classifier was used for the second layer to stay close to the setup by Custódio and Paraboni (2021). Overall, per target, the result of the experimental setup is comprised of 1200 SVMs as well stacked models, i.e., 2400 models in total for both targets. To enable a concise analysis, we opted to showcase the results for the text set for which individual inputs have a length of 500 characters. The results for the other text sets may be found in the Appendix.

Aggregate Overview for Feature-Stability

First, we will look at the aggregated results for our experiments. In Figure 5.1, we see the F1-scores for the classifiers trained on the dataset with a minimum character length of 500 per instance. As in previous results, the score declines markedly in the number of authors. However, especially the models trained either with different feature types as input (labeled ”cumulated”) or those trained similarly to the DynAA model by Custódio and Paraboni (2021) (labeled ”stacked”) perform consistently. Moreover, overall the results are in line with previous top-performing results on the PAN2019 dataset (Wiegmann et al., 2019). The baseline models trained on feature sets consisting of singular types have a high variance in performance. That is not surprising, when we look at individual models, each trained on one type of features. As can be seen in Table 5.2 and Table 5.3, the performance is on the upper end for the feature types such as CHAR with an $F1_{50}^{CHAR-5}$ up to 0.88/0.82 and on the lower end for feature types such as NUM with an $F1_{50}^{NUM}$ of 0.59/0.29 for the targets gender/age. Consequently, the feature types used, as well as their combinations, not only have a great impact on the outcome, but some features do encode little or next to no information for our classification task. Hence, we exclude those from our further analysis. The same findings apply to the models trained on instances with a minimum length of 100 characters and 250 characters respectively (Figure E.1 and Figure E.3 in the Appendix). Having comparable results in terms of accuracy and F1-score to what is found in the literature for this dataset serves as a basis for our following evaluation. A high performance lends credence to

[6] MPCDF HPC System "Raven".



(a) Results for target *gender*.

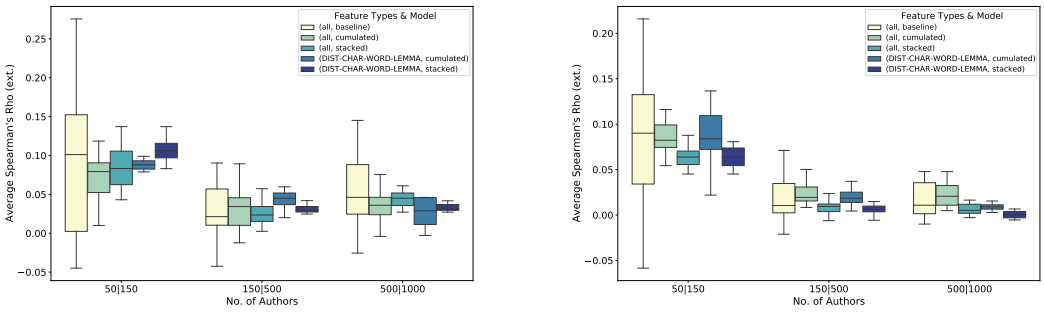
(b) Results for target *age*.

Notes: The figure shows boxplots for the F1-score of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure 5.1: F1-score input instance length of 500 characters.

the assumption that our classifier is indeed working well and extracting the relevant information from the input data. In reverse, we may therefore assume that the features used are indeed those holding the relevant information for the respective task. Thus, our approach of extracting the feature relevance via the associated weights is sensible.
Figure 5.2 shows the results for the extended distortion calculated via Spearman's Rho. The comparison is always done between two models, varying in the amount of authors the respective model is trained on. In the comparison, model 1 is trained on the lower number of authors (e.g., 50) whereas model 2 is trained on the next-highest number of authors (e.g., 150). Overall, we thus have 3 comparisons in the number of authors.
For target gender, on average, and irrespective of the model, we find only little correlation when increasing the number of authors from 50 to 150 ($0.05 < \rho < 0.20$). For the others, when increasing the number of authors, correlation goes down to a maximum of 0.05. At the same time, we show that the performance in terms of f1-score remains relatively stable. This implies that the stability in performance comes at the expense of the stability in feature relevance. When increasing the number of authors, different features are therefore predictive in terms of the target. These findings do not generally imply that the previous features lose their relevance. They do however mean that, when ordering all features in terms of the associated relevance, the ordering changes fundamentally. That fact is reflected in Spearman's Rho. Regardless by how much we increase the number of authors, on average the correlation lies between 5% and 15% for all models. That is interesting, as the number of features additionally available when increasing the number of authors is never above 30%. Thus, it cannot be that mainly those additional features are the ones being used for predictions, as the correlation coefficient would then still be higher than what we find. Rather, it seems to be the case that the model weighs the previous and new features in such a way that the new ordering imposed differs completely from

the previous one.



(a) Results for target *gender*.

(b) Results for target *age*.

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure 5.2: Extended Spearman correlation input instance length of 500 characters.

Similar results are found for the target age. However, here the decline in correlation is even more unidirectional when increasing the number of authors (an increase in authors yields a decline in correlation).

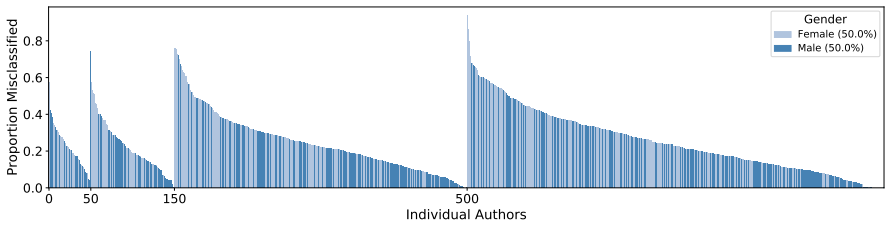
The only outlier to these results is the baseline model trained on individual feature types. While on average the correlation is the same as for the models on combinations of feature types, the outliers show a very high positive correlation (up to $\rho^{NUM}_{150|50} : 0.42$, see Table 5.3). While that may look like it stands in opposition to the other results, their low predictive power helps to explain this phenomenon. Table 5.3 shows that the F1-score for the prediction of age is only at $F1^{NUM}_{150} : 0.25$ with the random guess benchmark being 0.2. Thus, while the feature relevance remains stable for some features, their predictive power is negligible. Thus, the relevance assigned to these features may simply be random noise without any signal. As such, when looking at the average feature stability over all feature type sets, the conclusion is that the features are, on average, not stable and the distortion of relevance when increasing the number of authors in a dataset is already high for a low number of authors (from 50 to 150). These findings hold regardless of the character length of the input text, as Figure E.2 and Figure E.4 in the Appendix show.

Author-Level Analysis

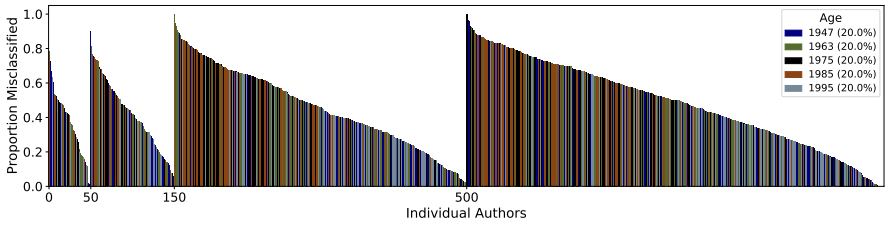
For the author-level analysis, we look at the results gained by feeding the classifier the full set of feature types in a cumulated way. That choice assures that our results are predicted making use of the full information. Furthermore, cumulating the input yields the best results overall (compare section E.2). In that way, the analysis is done using the best

possible set. However, the findings hold for any of the feature-type approach combinations.

Figure 5.3a shows the error at the author level when predicting gender. We see, that overall we have very few authors for which the accuracy is lower than the random-guess accuracy (0.5), i.e, for which our prediction error is higher than 0.5. We see that the relative number of authors for which the classifier performs below the random-guess threshold stays stable, when compared to the number of authors in the set. Consequently, the relative overall-classification performance at the author level stays stable, even when increasing the the number of authors in the set. However, there is a small but stable proportion of authors for which the classifier is systematically unable to predict the target correctly. When looking at the distribution of the errors across genders, we find that a



(a) Author-level errors for target *gender*.



(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

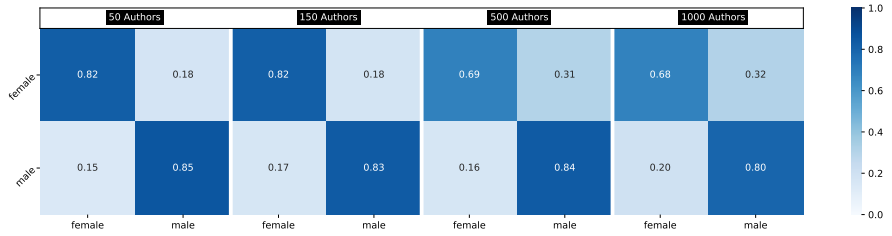
Figure 5.3: Author-level results for the full feature set with an input instance length of 500 characters.

higher number of those authors for whom our classifier makes systematic errors seems to be female. On the high level, we find that the performance remains relatively stable when we increase the number of authors as depicted in Figure 5.4a. However, as soon as we reach the two upper-most brackets of authors, we see that the result for male authors remains relatively stable, while the outcome for female authors declines markedly from $acc_{150}^{Female} : 0.82$ to $acc_{1000}^{Female} : 0.69$. For the same sets, the mostly stable accuracy for males declines only from $acc_{150}^{Male} : 0.83$ to $acc_{1000}^{Male} : 0.80$. Looking closer, we can see that

that drop for females corresponds to an increase in the number of authors from 150 to 500. The implication is thus that we add female authors who are systematically difficult to classify. Bringing this together with our observations from before, we include only those authors in Figure 5.4b for whom the classifier performs worse than the random-guess threshold. Here, we find some support for our previous assumption. When looking at the errors we make for female authors, we see that, starting from 500 authors onward, the number of those below the random-guess threshold jumps up.

While that seems to point to the fact that apparently female authors are difficult to classify, the true reason may be slightly more nuanced. When looking at the male authors in Figure 5.4b, we see that, while obviously low in absolute numbers, the pattern is inverse for the datsets comprised of 50 and 150 authors. The underlying driver, however, does not seem to be the gender per se, but rather the lack of stability in regards to feature importance. Per design, we limited the amount of features for each feature type to those appearing at least in 1% of all training instances. Hence, the number of features for word-based feature types increases only sublinearly relative to the number of authors. Thus, the amount of author-individual fitting the classifier is able to achieve declines. Indeed, that is the very essence of reducing overfitting. However, as shown in Figure 5.2a, the stability of feature importance is low. Taken together, that simply means that systematic patterns within a limited number of features for authors of the same gender decline when the number of authors increases, i.e., the patterns seem to be merely correlational – they start to break down or become unstable and more complex. The underlying reason is that additional authors introduce new features, while using the old features in a different way. As the classifier is only able to estimate one weight per feature and the number of additional features is limited, the ability to represent all the necessary information declines.

For the target age, we also find only comparatively few authors with an average accuracy below the random-guess threshold of 0.2 (see Figure 5.3b). When comparing it to the results of the target gender, it becomes clear that the number of those below the threshold jumps up significantly when the number of authors increases from 150 to 500. Moreover, it seems to be the case that authors of the intermediate age brackets (1975 and 1985) seem to be more difficult to classify. Overall, the patterns seem to be less pronounced when compared to the ones found for gender. Looking at the category-wise analysis presented in Figure 5.5a, we find that, overall there are only few pronounced patterns of confusion. As already suggested by Figure 5.3b, only the intermediate age brackets have a systematic pattern. Especially for the datasets consisting of 500 and 100 authors, the confusion between the true age bracket 1985 and the youngest age bracket 1995 is pronounced. Here, only 38% of the instances are classified correctly as 1985, while 28% are confused as 1995. While here the interpretation might be that the distinction between the youngest authors might be difficult, the results of 1975, the intermediate category, make it more difficult. We see that the confusion with the category 1963 as well as 1995 is of similar size. It might be that, instead of predicting only age, the classifier



(a) All authors (row-wise normalization).

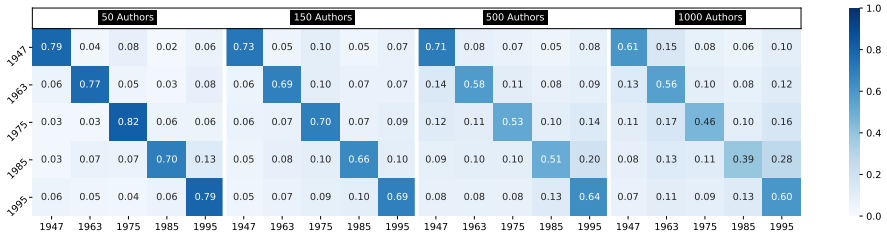
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 500 Characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

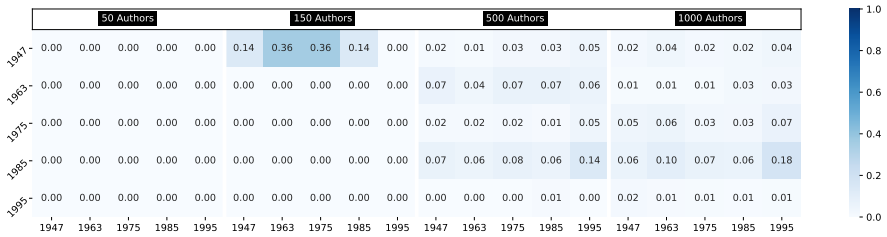
Figure 5.4: Confusion matrices for target *gender*.

picks up on a proxy in the way people express themselves. While certainly dependent on age in terms of punctuation for older authors (Flekova et al., 2016) as well as on stability of language use in younger authors (De Jonge and Kemp, 2012), the way of expression also depends on the groups to which we belong (Chan and Fyshe, 2018). Consequently, some of those authors confused might simply be part of peer groups where the mode of expression is reflective of younger age brackets. As a consequence, they get misclassified. These relatively distinctive patterns for the oldest and youngest authors are also most likely the reason why the prediction accuracy for those is markedly high, even for the set with the highest number of authors.

When looking only at those authors below the random-guess threshold, as depicted in Figure 5.5b, we find that there are only two age brackets for which we have a systematic and pronounced confusion. For the set with 150 authors, this is the age bracket 1947, which is most often confused with the two adjacent age brackets 1963 and 1975. For the set with 1000 authors, the age bracket 1985 is mostly confused with belonging either to the youngest or the oldest age bracket. As that bracket is also overall the most confused one, it stands to reason that the variance in expression is the highest. Consequently, there is no pronounced pattern in the features on which the classifier is able to pick up.



(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 500 Characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure 5.5: Confusion matrices for target *age*.

5.4 Robustness of Feature-Relevance

In order to assess how the aggregate results from the previous section hold, when looking at individual feature types as well as different classifier approaches, this section presents a more fine-grained insight. In this section, we focus on the input in terms of feature types and n-grams. We analyze these aspects with a special look on the stability of the feature relevance.

Baseline

First, we now take a look at the results in feature set 1, i.e., the results gained when using each feature type individually to predict the target. Table 5.2 and Table 5.3 show the experimental results for targets gender and age, respectively. Each row shows the result for a feature type and the corresponding n-grams. The feature types themselves are sorted in an ascending order such that feature types in lower rows capture more context. The type CHAR (characters), for example, captures, in principle, less contextual information (such as topic or structural information) compared to, for example, word-based n-grams (Rocha et al., 2017). Naturally, when the n-gram window is increased, e.g., for character-based features from 2 to 4, the character n-grams also start to capture contextual information. Consequently, the n-gram combinations within the individual

feature types are also sorted in an ascending fashion.

Table 5.2: F1-scores & stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST	2	0.6322	–	0.3481	0.0858	0.3680	0.1528	0.5403	0.2834
	2-3	0.6477	–	0.3788	0.1535	0.4365	0.1350	0.5173	0.1451
	2-3-4	0.6624	–	0.6056	0.1185	0.5462	0.0893	0.5432	0.0425
	2-3-4-5	0.6789	–	0.6402	0.0904	0.5385	0.0491	0.4858	0.0457
CHAR	2	0.8080	–	0.7580	0.2757	0.6882	0.0903	0.6719	0.0932
	2-3	0.8690	–	0.8151	0.1739	0.7638	0.0277	0.7260	0.0251
	2-3-4	0.8702	–	0.8411	0.1328	0.7746	0.0174	0.7362	0.0364
	2-3-4-5	0.8847	–	0.8094	0.0999	0.7756	0.0131	0.7411	0.0215
ASIS	2	0.8357	–	0.7995	0.2720	0.7184	0.0573	0.6916	0.0588
	2-3	0.8868	–	0.8436	0.1492	0.7750	0.0155	0.7341	0.0213
	2-3-4	0.9004	–	0.8662	0.1229	0.7885	0.0114	0.7493	0.0150
	2-3-4-5	0.9040	–	0.8158	0.1025	0.7882	0.0207	0.7488	0.0153
POS	1	0.5907	–	0.5847	0.2456	0.5775	0.3123	0.5728	0.1982
	1-2	0.6303	–	0.6403	0.0272	0.6061	-0.1515	0.5985	0.2639
	1-2-3	0.6583	–	0.6579	-0.0449	0.6174	-0.0222	0.6127	0.0582
TAG	1	0.6402	–	0.5700	-0.0086	0.5892	-0.1640	0.5771	0.0043
	1-2	0.6982	–	0.6704	0.1546	0.6265	0.0233	0.6140	0.1140
	1-2-3	0.6980	–	0.6828	0.0514	0.6230	0.0017	0.6278	0.1243
DEP	1	0.6195	–	0.5997	0.0103	0.5652	-0.0425	0.5655	0.0244
	1-2	0.6590	–	0.6277	0.1047	0.5963	-0.1028	0.5926	-0.0255
	1-2-3	0.6704	–	0.6550	-0.0349	0.6008	0.0236	0.6060	0.0737
LEMMA	1	0.7786	–	0.7679	-0.0374	0.7096	0.0221	0.6869	0.0406
	1-2	0.8007	–	0.7816	-0.0143	0.7144	0.0099	0.6925	0.0467
WORD	1	0.7588	–	0.7408	-0.0304	0.6940	0.0136	0.6782	0.0480
	1-2	0.7653	–	0.7535	-0.0003	0.6981	0.0560	0.6836	0.0304
NUM	1	0.5912	–	0.5635	0.2500	0.5229	0.0833	0.4791	0.0667

Looking at results for the target gender, we first find that the most predictive feature types are those most closely related to the words of the text, but not necessarily the structure. That can be seen from the fact that structure-capturing feature types, such as POS, TAG, and DEP, show low predictive power no matter what the number of authors within the subset is. Moreover, CHAR-2-grams already perform well ($F1_{50}^{CHAR-2} : 0.80$) on the small dataset comprised of 50 authors. However, when we increase the number of authors, the performance declines markedly ($F1_{1000}^{CHAR-2} : 0.67$), especially when compared to CHAR-(2,5)-grams ($F1_{1000}^{CHAR-5} : 0.741$), which are close to the top performance ($F1_{1000}^{ASIS-5} : 0.748$). The same pattern, although on a lower overall performance level, is visible for the text distortion features DIST capturing punctuation and other stylistic markers. For lower n-gram sizes, the performance is only negligibly above or below the random guess threshold, while for higher n-grams the performance is higher ($F1_{50}^{DIST-5} : 0.67$), but then decreases again in the number of authors. Consequently, the results show that there seems little cause to think that there are patterns in the style of authors related to gender. On the other side, CHAR-2-grams have a reliable performance ($0.67 < F1^{CHAR-2} < 0.80$); increasing the n-gram window only by 1 increases performance even more. Consequently, it can be assumed that the there seems to be a discernible pattern related to gender within the character combinations used. The

underlying assumption would be that certain topics might be reflected by the use of similar words or that certain synonyms are preferred by one group over the other. We can compare this with the result for the WORD-grams. Here we see that, while we the performance is high, it is still worse when compared to CHAR-(2,5)-grams. The latter would also capture words up to five characters long. However, if that overlap is the sole driver of performance, then WORD-grams should not be outperformed. As such, we can conclude that there is a discernible pattern related to gender in low-context CHAR-n-grams. In terms of the stability of the feature relevance, the results are sobering. As in the aggregate before, the correlation tends towards zero when increasing the number of authors. Besides that, in some cases the correlation even flips signs. That implies features which were useful for predicting group A before are now either relevant for neither group or relevant for predicting group B (see, for example, Table 5.2, $\rho_{500}^{POS-2} : -0.15$). While mostly small, all correlation coefficients are significant at the 1%-level.
When looking at age, the results shown in Table 5.3 reflect the overall findings for gender. Text distortion alone, such as punctuation reflected in the features of type DIST, does hold some, but not the majority of the information relevant to the prediction of age. That is evident from the stark decline towards random-guess accuracy, especially for low-level n-grams.

Table 5.3: F1-scores & stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Age						
	Min. No. of Characters	500						
	No. of Authors	50		150		500	1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST	2	0.4797	–	0.2249	0.1872	0.2808	0.0083	0.2196	0.0192
	2-3	0.4809	–	0.3711	0.1142	0.2633	0.0275	0.2736	0.0478
	2-3-4	0.5294	–	0.4482	0.0268	0.2454	0.0502	0.2921	0.0099
	2-3-4-5	0.5384	–	0.4088	0.0219	0.2401	0.0371	0.2823	0.0154
CHAR	2	0.7257	–	0.6250	0.2159	0.4761	0.0197	0.4033	0.0058
	2-3	0.8024	–	0.7319	0.1388	0.5696	0.0014	0.5070	-0.0014
	2-3-4	0.8274	–	0.7579	0.1140	0.5938	0.0124	0.5346	0.0123
	2-3-4-5	0.8291	–	0.7492	0.0916	0.6114	0.0075	0.5407	0.0068
ASIS	2	0.7811	–	0.6776	0.2031	0.5142	0.0049	0.4426	-0.0024
	2-3	0.8546	–	0.7646	0.1386	0.6048	0.0027	0.5346	0.0025
	2-3-4	0.8687	–	0.7836	0.1080	0.6189	0.0032	0.5512	0.0069
	2-3-4-5	0.8542	–	0.7890	0.0925	0.6174	0.0023	0.5523	0.0011
POS	1	0.3529	–	0.2723	0.1807	0.0964	0.1046	0.1010	0.1923
	1-2	0.4794	–	0.3512	0.0762	0.2269	-0.0211	0.2283	0.0116
	1-2-3	0.5248	–	0.4124	0.0611	0.2767	0.0236	0.2491	0.0455
TAG	1	0.3934	–	0.3244	0.0886	0.1338	0.0711	0.1690	-0.0680
	1-2	0.5432	–	0.4479	0.0178	0.2755	-0.0109	0.2670	0.0441
	1-2-3	0.5878	–	0.4859	0.0825	0.3811	0.0386	0.3110	0.0474
DEP	1	0.3560	–	0.2418	0.0971	0.1000	0.0466	0.2387	0.0411
	1-2	0.5057	–	0.3820	0.0559	0.2523	0.0155	0.2358	-0.0100
	1-2-3	0.5369	–	0.4029	0.0751	0.3292	-0.0095	0.2873	-0.0020
LEMMA	1	0.6668	–	0.5763	-0.0219	0.4091	0.0128	0.3339	0.0078
	1-2	0.6920	–	0.5881	0.0010	0.4424	0.0085	0.3781	-0.0058
WORD	1	0.6282	–	0.5416	-0.0585	0.3867	0.0428	0.3439	0.0151
	1-2	0.6437	–	0.5482	-0.0112	0.4053	0.0013	0.3294	0.0118
NUM	1	0.2995	–	0.2542	0.4267	0.2386	-0.0933	0.2194	0.3033

When combined with CHAR, then especially higher-order n-grams (which is reflected in the feature type ASIS) hold the most information about an author's age. That seems

to be in line with findings linking age to a higher adherence to linguistic rules, even in an online environment (De Jonge and Kemp, 2012; Hovy and Søgaard, 2015). However, the content of the tweets also seems to set the age categories apart, as illustrated by the fact that TAG alone has a relatively high predictive power even for the dataset comprised of 1000 authors (F1: 0.31). The same holds true for LEMMA, implying that age groups are also set apart by the use of one set of words over another. Here again, the feature stability is low, with a $\rho \in [0, 0.05]$.

Thus, we can conclude that, for singular feature sets, the model is able to extract information from the features, especially those with higher context, as evident from the increase in predictive performance when the n-gram range is increased. However, the relevant information is not stable in the number of authors, which means that additional authors introduce a wider variation, that needs to be separated differently than the smaller range. As the number of characters is limited overall (and thus the number of features in the lower n-gram range), that automatically implies that the content and therefore the relevant features change. That seems to lead to an overall change in the way individual features are predictive. Thus, the rank correlation is low.

Table 5.4: F1-scores & stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST_CHAR	2	0.8230	–	0.7868	0.0911	0.7058	0.0598	0.6769	0.0998
	2-3	0.8802	–	0.8392	0.0903	0.7696	0.0331	0.7311	0.0178
	2-3-4	0.8712	–	0.8494	0.0990	0.7852	0.0445	0.7392	0.0454
	2-3-4-5	0.8951	–	0.8583	0.0829	0.7650	0.0455	0.7429	0.0461
DIST_CHAR_ASIS	2	0.8882	–	0.8623	0.0815	0.7441	0.0306	0.7494	0.0118
	2-3	0.8850	–	0.8725	0.0931	0.8036	0.0462	0.7602	0.0295
	2-3-4	0.8764	–	0.8733	0.0789	0.7641	0.0380	0.7675	0.0026
	2-3-4-5	0.8942	–	0.8729	0.0829	0.7626	0.0459	0.7673	0.0281
DIST_CHAR_ASIS_LEMMA	1	0.8925	–	0.8697	0.0935	0.7765	0.0201	0.7706	0.0152
	1-2	0.8779	–	0.8744	0.0829	0.7740	0.0428	0.7703	0.0084
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8797	–	0.8712	0.0632	0.8034	0.0388	0.7688	-0.0029
	1-2	0.8943	–	0.8702	0.0526	0.7702	0.0268	0.7723	0.0094

Cumulated

The previous analysis has shown that some feature types, such as POS, TAG, DEP, and NUM do hold little relevant information. We therefore constructed a subset of feature types which excludes them. That subset includes the types ASIS, CHAR, LEMMA, and WORD. Compared to the previous analysis, we now give the model the possibility to include additional information, i.e., information stemming from different feature types, in the model. As shown by the results in Table 5.4 and Table 5.5, the additional information yields an overall increase in performance. How much additional information leads to an improvement differs by target. For age, we find that including only little additional contextual information already increases the outcome. However, when the contextual

information becomes larger, e.g., by including LEMMA and WORD, the result does not improve anymore. The result is consistent across a different number of authors. Consequently, the information for age seems to be less reliant on contextual information and content. Already single-word content and context as captured by CHAR-(2,5) and ASIS-(2,5), is enough for a high prediction score. When we compare the outcome for gender with the the results in Table 5.2, we see that using a cumulated input improves the results overall. It is especially important to note that, when faced with a high number of individual authors, increasing the context by using additional feature types such as LEMMA or WORD in addition to high-level n-grams increases performance. When taken together, our findings show that context and underlying data structure is an important driver behind the predictions of a model, as shown by the fact that the relevant features in terms of predictiveness change. At the same time, we show that the weight placed on individual features (and thus individual inputs reflecting certain contexts) is not stable. That is evident by the correlation scores across different author

Table 5.5: F1-scores & stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)	F1-score	Avg. Spearman's ρ (ext.)
Feature types	N-gram ranges								
DIST_CHAR	2	0.7271	–	0.6349	0.1366	0.5123	0.0244	0.4434	0.0048
	2-3	0.8214	–	0.7331	0.1364	0.5898	0.0286	0.5177	0.0117
	2-3-4	0.8343	–	0.7627	0.1078	0.6114	0.0183	0.5390	0.0110
	2-3-4-5	0.8250	–	0.7595	0.0934	0.6140	0.0179	0.5397	0.0084
DIST_CHAR_ASIS	2	0.8518	–	0.7727	0.0901	0.6150	0.0189	0.5515	0.0071
	2-3	0.8726	–	0.7954	0.0815	0.6629	0.0157	0.5687	0.0102
	2-3-4	0.8753	–	0.8014	0.0848	0.6610	0.0124	0.5553	0.0053
	2-3-4-5	0.8671	–	0.8006	0.0831	0.6376	0.0198	0.5731	0.0067
DIST_CHAR_ASIS_LEMMA	1	0.8608	–	0.7966	0.0753	0.6478	0.0143	0.5717	0.0077
	1-2	0.8615	–	0.7963	0.0736	0.6629	0.0197	0.5544	0.0028
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8515	–	0.7883	0.0680	0.6558	0.0044	0.5736	0.0056
	1-2	0.8293	–	0.7727	0.0682	0.6652	0.0116	0.5758	0.0095

sets. The scores are $\rho^{NUM}_{150|50} : 0.42$ at the highest for target gender and $\rho^{DEP-2}_{150|50} : -0.1$ at their lowest. Consequently, while the predictive accuracy is high, the model seems to rely on correlational patterns which are not only not invariant, but also quite unstable when the dataset is changed only slightly.[7]

5.5 Discussion

Overall, we find that the classifier makes systematic errors at the author level. For the target gender, one group of authors (female) seems to be difficult to classify in general. The underlying drivers seem to be that, for the shown dataset, the group as such has a very heterogeneous pattern in the features. In other words, the second group (male)

[7]All analyses –stacked as well as cumulated – were also performed on the full number of feature types, as well as the different numbers of authors and the different input instance lengths. The results may be found in the section E.2.

seems to be simpler to classify. That, however, may also be driven by the fact that the context in which they are active is more homogeneous than for female authors. For age, we find slightly more stable results, as most age brackets exhibit clear patterns that make the individual age brackets distinguishable. In terms of stability, however, the results are more mixed. While the features mainly driving the prediction are not context-reliant per-se, increasing the available context does increase performance markedly. This is especially evident from the fact that a wide n-gram window for character features yields the greatest relative increase in performance, outperforming even those models for which additional context information is made available by including additional feature types known to capture context. That in itself is not directly surprising and not necessarily cause for any concern. However, it is important to note that, for a low number of authors, about 10% of the prediction performance stems from an increase in context (e.g. for features CHAR and ASIS when predicting gender, see Table 5.2). When looking at the dataset with the largest number of authors, additional information is also what makes the model perform slightly better than the random-guess threshold and pushes it into the performance ranges found within the literature for comparable data (Wiegmann et al., 2019). Moreover, building a model on top of a composite of feature types (or stacking it on top; see Table 5.4 and Table 5.5) is when we see additional performance increases, especially for longer input texts. Thus, giving additional context on top of non-context feature types yields a better decision boundary for the classifier.

These results at first seem like technical details. However, in practice they show that the context the model is trained on and in (as simulated by varying the number of authors) largely carries over into its predictive performance. That means models trained within one context may not simply be used in another one. That is intuitive. What we show here, however, is that even by staying within one group of individuals (creators) and within one domain (Twitter), an increase in the number of possible targets changes the relevant features, and it also significantly changes the information as well as the context encoded within . That becomes evident from the fact that the stability in the relevance of features simply does not exist. For social sciences, these findings are relevant on two fronts. First, the models using the features presented here are indeed well-suited to find a pattern connecting their use to the prediction target. However, that pattern is unstable, changing with the number of authors or features available. Consequently, it hints at the fact that these patterns are merely correlations exploited by the model. Such correlations are difficult to rely upon, as their patterns – as shown by increasing the number of authors – may change at any time. Thus, this calls for a careful assessment of the validity when employing pre-trained models within the field, especially when the prediction outcome is used as input for further models or for further analysis. In other words, a change in behavior by individuals – either over time or by choice – will render the learned context irrelevant. Thus, the environment during training must be carefully compared to the one in which the model is used. In general, the findings thus paint a bleak picture for the social sciences. Our results show that there are authors of certain groups for which one has to expect above-average errors and systematic patterns of

misclassification. Taken together with the apparent lack of stability in the predictiveness of features when the dataset changes slightly means that, even when these patterns are assessed during training, the researcher has little chance to assess how they will affect a the result during the time of use. The differences between training data and test data might be difficult to pinpoint. Thus, for cases where it is not clear by how much training context and use-context differ, a social scientist should be very careful in simply adopting pre-trained models as the size of the introduced error is unknown.

Another finding is of a normative nature and tied to the wider debate of transparency and proportionality, and thus affects in particular the field of law. As law enforcement is faced with the problem of combing through a large amount of online content, searching and assessing such content by hand is untenable. Thus, already today algorithms are employed by law enforcement. However, especially in such environments, it must be clear how much of the findings by an algorithm is relies merely on correlations and especially how stable these correlations are in different environments. That does not even include the fact that there might be some groups of individuals for whom the classifier makes systematic mistakes. Only then do law enforcement, the defendant, and also the courts have the possibility to assess the validity of an result before acting upon it. After all, how valid is a result identifying traits of a suspect when the features are context-reliant to such a high degree that changing the use of some emojis or some words would alter the result completely? How robust is a result, when the features driving the result change with the number of authors an individual is compared to? What is even more problematic in real-world terms is that the instances used for training and those used for during actual application are separated from each other by time. Thus, a real culprit could evade being identified simply because the context changes, while innocents could be systematically misidentified as culprits. Thus, as the stability in feature relevance is already lacking for the relatively small changes introduced here, we should ask ourselves what requirements an algorithm should fulfill before it is being used within the law enforcement context. Optimally, we would ask for causal relationships between input and output. However, that might not be possible. The second-best would then be to have transparency for the model and a some-what stable relationship between input and output. The former assures that users, i.e., the state, as well as affected individuals are able to assess the inner workings of a model. That would enable an individual to judge whether the prediction pertaining to them might be part of a systematic error.

A reasonably stable relationship between input and output, i.e., a stable feature relevance, guarantees that while there might be systematic errors, the affected groups stay at least constant, although the dateset for the predictions may vary slightly compared to the training dataset, e.g., by number of authors or point-in-time. The alternative would be, of course, to specify a "half-life" before a model has to be re-trained and re-assessed.

As the findings of this study point towards such an unstable relationship, we argue that the features used in tasks related to authorship profiling and authorship attribution need

much more research. Moreover, models should be assessed with a measure for defining the boundaries of their stability. The result of that measure has to be affixed to the model so users may be able to infer its usability. Otherwise, establishing a scientifically valid link – going beyond merely showing that the model yields good correlational predictions on some datasets – might be impossible.

APPENDIX – CHAPTER 1

A.1 Extensive Analysis

Table A.1: Wilcoxon rank sum test with continuity correction - Likelihood Vignette with the Categories Present - Future

	Data Likelihood Vignettes on Tense Framing			
	Federal Audit Office	McKinsey	Buxtehude	Chopin
	(1)	(2)	(3)	(4)
W	148940	145120	160740	149800
p-value	0.023	0.003	0.89	0.034
Note:	alternative hypothesis: true location shift is not equal to 0			

Table A.2: $\tilde{\chi}^2$-test Immediacy Vignettes with the Categories Present - Future

	Data Immediacy Vignettes on Tense Framing			
	BonnEconLab	Halle/Saale	Mansfeld	Bitcoin
$\tilde{\chi}^2$-test	14.25	88.56	19.17	8.86
df	10	10	10	10
p-value	0.16	$1.032 * 10^{-16}$	0.038	0.54

Note: Alternative hypothesis: "Estimated point in time and 'framing language' are not independent."

A.2 Design of the Experimental Tasks

Time Preference Elicitation Task

Table A.3: Illustration of the Payment Schemes

	Payment in week after experiment in €														
Schedule	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	27.95	27.95	27.95	27.95	27.95	27.95									
2		28.64	28.64	28.64	28.64	28.64	28.64								
3			29.26	29.26	29.26	29.26	29.26	29.26							
4				29.8	29.8	29.8	29.8	29.8	29.8						
5					30.26	30.26	30.26	30.26	30.26	30.26					
6						30.64	30.64	30.64	30.64	30.64	30.64				
7							30.95	30.95	30.95	30.95	30.95	30.95			
8								31.18	31.18	31.18	31.18	31.18	31.18		
9									31.33	31.33	31.33	31.33	31.33	31.33	
10										31.4	31.4	31.4	31.4	31.4	31.4

The purpose of the time elicitation task in experiment 1 is to classify subjects by their personal discount rate. In this task, the effect of linguistic framing (tense) on time preferences was analyzed. Hence, it was sought to design a task in such a way that subjects could choose payment schedules. This choice should then automatically imply different personal discount rates.
Assumptions and Constraints:

1. The range of relevant personal discount rates was thought to lie between 2.25% and 36.25%. Personal discount rates within this band seem to be the most common in laboratory experiments for subject groups similar to the one participating in experiment 1 and frequently-cited studies regarding time preferences supported this assumption (Harrison et al., 2005). The range was then divided up into 10 intervals as a compromise between attempted fine-graininess and not wanting to overwhelm participants of the experiment with a multitude of choices.

2. The discounting function of subjects can be modelled by a utility function that is both increasing and concave. For this reason, the log-utility specification, i.e., $u(x) = ln(x)$, was chosen. This utility specification is often used in applications which require a specification.

3. For the final schedules, the midpoint of each interval was chosen to calculate the weekly amounts.

4. The weekly amounts were calculated transforming the yearly discount rate into a weekly one.

Payment Schedule:
Each schedule ran for six weeks and consisted of a weekly payment, the amount of which

was fixed.[1] Each schedule had a different starting date for the first payment. This simulates different multiple-choice lists for which we only showed an optimal switching point for one particular discounting factor. The payment schedules were designed as follows:

1. Schedule 1: This is the schedule targeting the most impatient group of subjects, i.e., subjects with a personal discount rate in the interval of $(0.3625; 0.3175)$.
2. Schedule 1 pays an amount S_1 every week, starting 1 week after the experiment and lasting six weeks in total.
3. The schedule targeting the second-most impatient group of subjects, i.e., those with a personal discount rate in the interval of $(0.3175; 0.2825)$, is called schedule 2 and so forth.
4. Schedule 2 pays an amount S_2 every week, starting 2 weeks after the experiment for six weeks in total.
5. Schedule 3 pays an amount S_3 every week, starting 3 weeks after the experiment and for six weeks in total and so forth.

For a graphical illustration of the payment schedules, please see the Table A.3.

Calculation of the Optimal Schedules for Subject Groups:
Each bracket of time preferences corresponds to a certain group of subjects. Thus, they are identified by revealing their preferences through their choice of a payment schedule. It is important that the payment schedules calculated are optimal for the different subject groups, i.e., schedule 1 should be optimal for the most impatient subject group and so forth. Schedule 10 is the optimal schedule for the most patient subjects and therefore must yield the most money on a weekly basis, i.e., $S_{10} = 31.4$ Euro.
From the optimality conditions of payment schedules, it follows that:

$$u(schedule10|r \in (0.025; 0.0475)) \geq u(allotherschedules|r \in (0.025; 0.0475)) \quad \text{(A.1)}$$

,in particular

$$u(schedule10|r = 0.0475) \geq u(schedule9|r = 0.0475) \quad \text{(A.2)}$$

The same optimality conditions also need to hold for subjects who are targeted by schedule 9. Hence, the following needs to hold.

$$u(schedule9|r \in (0.0475; 0.0825)) \geq u(allotherschedules|r \in (0.0475; 0.0825)) \quad \text{(A.3)}$$

,in particular

$$u(schedule9|r = 0.0475) \geq u(schedule10|r = 0.0475) \quad \text{(A.4)}$$

[1] If we amended the amounts on a weekly basis to factor in the additional change induced by a weekly discount rate, the amounts would only change at the second, third, or even fourth decimal.

and

$$u(schedule9|r = 0.0825) \geq u(schedule8|r = 0.0825) \tag{A.5}$$

From the optimality of both schedule 10 and schedule 9 at $r = 0.0475$, it is possible to deduce the following equality :

$$u(schedule10|r = 0.0475) = u(schedule9|r = 0.0475) \tag{A.6}$$

Equation A.6 uniquely determines the payment for S_9 as a function of the initial payment of S_{10}. The remaining values S_8 as well as the others can the be calculated recursively, i.e.:

$$u(schedule8|r \in (0.0825; 0.1175)) \geq u(allotherschedules|r \in (0.0825; 0.1175)) \tag{A.7}$$

,in particular

$$u(schedule8|r = 0.0825) \geq u(schedule9|r = 0.0825) \tag{A.8}$$

and

$$u(schedule8|r = 0.1175) \geq u(schedule7|r = 0.1175) \tag{A.9}$$

This determines S_8 uniquely as a function of S_9 (and therefore of S_{10}). By iteration, this procedure allows for the identification of all $S_x \forall x \in (1, 2, ..., 9)$ from the externally given starting point S_{10}.

Table A.4: Ordered Logit Estimations for Time Preference Task (shortened)

	(1)	(2)	(3)	(4)	(5)	(6)
PF	−0.023	0.065	−0.226	−0.214	−0.181	−0.065
	(0.112)	(0.162)	(0.225)	(0.228)	(0.232)	(0.234)
TR[a]		0.261	0.265	0.294	0.281	0.348*
		(0.160)	(0.160)	(0.162)	(0.164)	(0.166)
Present-Pref[b]			−0.068	−0.073	−0.067	−0.045
			(0.073)	(0.073)	(0.075)	(0.075)
PF|TR[a]		−0.180	−0.172	−0.181	−0.193	−0.303
		(0.225)	(0.226)	(0.229)	(0.231)	(0.235)
PF|Present-Pref[b]			0.188	0.198	0.178	0.158
			(0.103)	(0.105)	(0.107)	(0.108)
Add. Controls[c]:	No	No	No	Yes	Yes	Yes
			Obs.: 1,137			

Note: Ordered Logit estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the blockwise order of time-preference and risk-preference task and PF framing. Order Time → Risk = Likelihood = 1.

b Number of sentences in the present tense ({0...5}) selected during paragraph construction task.

c Additional controls include: gender, alcohol consumption (beer/wine/spirits, smoking, sports per week, marital status, avalable monthly income, language-related studies, pet owner).

Table A.5: Ordered Logit Estimations for Time Preference Task

	(1)	(2)	(3)	(4)	(5)	(6)
PF	−0.023	0.065	−0.226	−0.214	−0.181	−0.065
	(0.112)	(0.162)	(0.225)	(0.228)	(0.232)	(0.234)
IL		0.261	0.265	0.294	0.281	0.348*
		(0.160)	(0.160)	(0.162)	(0.164)	(0.166)
Present-Pref			−0.068	−0.073	−0.067	−0.045
			(0.073)	(0.073)	(0.075)	(0.075)
Self Estimation Risk (0 = fully risk averse, 11 = risk loving)				0.018	0.017	0.006
				(0.031)	(0.031)	(0.032)
Beer:occasionally				−0.006	−0.082	−0.079
				(0.184)	(0.188)	(0.194)
Beer:seldomly				0.139	0.049	−0.013
				(0.202)	(0.206)	(0.217)
Beer:never				−0.153	−0.311	−0.310
				(0.236)	(0.243)	(0.258)
Wine:occasionally				0.101	0.081	0.143
				(0.240)	(0.243)	(0.247)
Wine:seldomly				0.035	0.012	0.119
				(0.245)	(0.247)	(0.255)
Wine:never				0.115	0.058	0.140
				(0.268)	(0.272)	(0.283)
Spirit:occasionally				1.062*	1.061*	1.117**
				(0.421)	(0.423)	(0.429)
Spirit:seldomly				0.921*	0.901*	0.952*
				(0.421)	(0.425)	(0.430)
Spirit:never				0.926*	0.873	0.897
				(0.449)	(0.452)	(0.459)
Mixed-Drinks:occasionally				−0.221	−0.105	−0.217
				(0.462)	(0.461)	(0.468)
Mixed-Drinks:seldomly				−0.243	−0.094	−0.188
				(0.457)	(0.457)	(0.466)
Mixed-Drinks:never				−0.220	−0.056	−0.171
				(0.473)	(0.472)	(0.486)
Smoker (1 = Yes)				0.596**	0.573**	0.498*
				(0.201)	(0.203)	(0.206)
Sport per Week (1 = more than 3 times a week, 4 = never)				0.025	0.018	−0.018
				(0.067)	(0.068)	(0.069)
Healthy Food:much heed				0.047	0.030	0.053
				(0.209)	(0.213)	(0.215)
Healthy Food:little heed				0.167	0.145	0.151
				(0.210)	(0.213)	(0.216)
Healthy Food:no heed				0.142	0.102	0.149
				(0.386)	(0.396)	(0.408)
Academic Degree Parents (1 = Yes)					−0.451***	−0.377**
					(0.118)	(0.120)
Marital St. Parents:married					−0.019	−0.102
					(0.315)	(0.322)
Marital St. Parents:divorced					−0.232	−0.413
					(0.334)	(0.343)
Age of first exposure to second tongue					−0.040*	−0.034
					(0.017)	(0.018)
Monthly available:between € 150 and < € 200					0.129	0.126
					(0.174)	(0.177)
Monthly available:between € 200 and <€ 250					−0.011	0.042
					(0.206)	(0.209)
Monthly available:between € 250 and <€ 300					−0.324	−0.273
					(0.233)	(0.237)
Monthly available:between € 300 and <€ 350					−0.027	0.039
					(0.240)	(0.244)
Monthly available:between € 350 and <€ 400					−0.260	−0.282
					(0.260)	(0.264)
Monthly available:between € 400 and <€ 450					−0.492	−0.484
					(0.293)	(0.295)
Monthly available:between € 450 and <€ 500					−0.172	−0.024
					(0.367)	(0.376)
Monthly available:between € 500 and <€ 550					−0.395	−0.319
					(0.365)	(0.375)
Monthly available:between € 550 and <€ 600					−0.388	−0.283
					(0.441)	(0.451)
Monthly available:between € 600 and <€ 650					−0.082	−0.026
					(0.498)	(0.511)
Monthly available: > € 650					−0.254	−0.271
					(0.258)	(0.272)
Age						−0.001
						(0.015)
Gender:female						−0.016
						(0.143)
Gender: NA						−0.463
						(0.484)
Most recent math grade						0.298***
						(0.068)
Marital status:married						0.048
						(0.331)
Marital Status:divorced						0.748
						(0.687)
Lingual-focused Studies:True						0.344*
						(0.138)
Pet:True						0.342**
						(0.127)
PF|TR		−0.180	−0.172	−0.181	−0.193	−0.303
		(0.225)	(0.226)	(0.229)	(0.231)	(0.235)
PF|Present-Pref			0.188	0.198	0.178	0.158
			(0.103)	(0.105)	(0.107)	(0.108)
1|2	−2.107***	−2.239***	−2.154***	−3.277***	−2.579***	−3.420***
	(0.111)	(0.139)	(0.166)	(0.616)	(0.712)	(0.793)
2|3	−1.849***	−1.982***	−1.896***	−3.016***	−2.316**	−3.154***
	(0.104)	(0.133)	(0.161)	(0.615)	(0.711)	(0.791)
3|4	−1.521***	−1.653***	−1.567***	−2.682***	−1.981**	−2.814***
	(0.096)	(0.126)	(0.157)	(0.613)	(0.710)	(0.790)
4|5	−1.214***	−1.346***	−1.260***	−2.372***	−1.669*	−2.495**
	(0.091)	(0.122)	(0.153)	(0.613)	(0.709)	(0.790)
5|6	−0.887***	−1.018***	−0.932***	−2.039***	−1.331	−2.150**
	(0.086)	(0.119)	(0.150)	(0.612)	(0.709)	(0.789)
6|7	−0.685***	−0.816***	−0.730***	−1.834**	−1.121	−1.934*
	(0.084)	(0.117)	(0.149)	(0.611)	(0.709)	(0.789)
7|8	−0.426***	−0.557***	−0.469**	−1.570*	−0.849	−1.655*
	(0.083)	(0.116)	(0.148)	(0.611)	(0.709)	(0.788)
8|9	−0.266**	−0.396***	−0.309*	−1.407*	−0.680	−1.481
	(0.082)	(0.115)	(0.148)	(0.611)	(0.709)	(0.788)
9|10	−0.177*	−0.307**	−0.219	−1.316*	−0.586	−1.384
	(0.082)	(0.115)	(0.148)	(0.610)	(0.709)	(0.788)
Obs.:	1,137					

Note: OLS estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.

b Number of sentences in the present tense ({0…5}) selected during paragraph construction task.

c The baseline category for the variables “beer", “wine", “spirit", and “mixed-drinks" is often.

d The baseline category for monthly available is "<€ 150".

e The baseline category for “gender" is “male".

f The values for the variable “math grade" spans the interval [1,5] with “1" being the best.

f The values for the variable "math grade" spans the interval [1,5] with "1" being the best.

Table A.6: OLS Estimations for Risk Preference Task (shortened)

	(1)	(2)	(3)	(4)	(5)	(6)
PF	0.973	−0.337	−0.261	−0.666	−0.530	−0.454
	(1.123)	(1.627)	(2.167)	(2.158)	(2.198)	(2.173)
TR[a]		−2.458	−2.438	−2.755	−2.929	−2.974
		(1.608)	(1.613)	(1.600)	(1.584)	(1.552)
Present-Pref[b]			−0.283	−0.188	−0.157	−0.259
			(0.724)	(0.719)	(0.721)	(0.719)
PF|TR[a]		2.646	2.611	3.282	3.472	3.335
		(2.243)	(2.248)	(2.222)	(2.231)	(2.204)
PF|Present-Pref[b]			0.016	−0.057	−0.106	0.091
			(0.996)	(0.988)	(1.001)	(0.990)
Add. Controls[c]:	No	No	No	Yes	Yes	Yes
			Obs.: 1,137			

Note: OLS. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of time-preference and risk-preference task and PF framing. Order Time → Risk = Likelihood = 1.

b Number sentences in the present tense ($\{0 \ldots 5\}$) selected during paragraph construction task.

c Additional controls include: gender, alcohol consumption (beer/wine/spirits, smoking, sports per week, marital status, avalable monthly income, language-related studies, pet owner).

Table A.7: OLS Estimations for Risk Preference Task

	(1)	(2)	(3)	(4)	(5)	(6)
PF	0.973	−0.337	−0.261	−0.666	−0.530	−0.454
	(1.123)	(1.627)	(2.167)	(2.158)	(2.198)	(2.173)
IL		−2.458	−2.438	−2.755	−2.929	−2.974
		(1.608)	(1.613)	(1.600)	(1.584)	(1.552)
Present-Pref			−0.283	−0.188	−0.157	−0.259
			(0.724)	(0.719)	(0.721)	(0.719)
Self Estimation Risk (0 = fully risk averse, 11 = risk loving)				1.090***	1.130	1.048***
				(0.323)	(0.322)	(0.324)
Beer:occasionally				−5.768***	−5.771	−3.939***
				(1.671)	(1.685)	(1.730)
Beer:seldomly				−8.236***	−7.902	−5.115***
				(1.908)	(1.934)	(2.021)
Beer:never				−4.884*	−4.402	−0.772*
				(2.367)	(2.366)	(2.460)
Wine:occasionally				−4.211	−4.189	−4.949
				(2.317)	(2.327)	(2.370)
Wine:seldomly				−2.335	−2.512	−4.483
				(2.391)	(2.375)	(2.449)
Wine:never				−0.034	0.106	−2.396
				(2.581)	(2.600)	(2.690)
Spirit:occasionally				7.837*	7.622	7.851*
				(3.499)	(3.530)	(3.548)
Spirit:seldomly				7.131*	7.191	8.313*
				(3.474)	(3.498)	(3.520)
Spirit:never				7.289	7.586	9.179
				(3.774)	(3.802)	(3.818)
Mixed-Drinks:occasionally				3.611	3.071	1.312
				(4.177)	(4.155)	(4.059)
Mixed-Drinks:seldomly				5.032	4.422	1.887
				(4.116)	(4.092)	(4.010)
Mixed-Drinks:never				5.192	4.381	0.732
				(4.339)	(4.302)	(4.248)
Smoker (1 = Yes)				−6.692***	−6.489	−5.777***
				(1.819)	(1.840)	(1.839)
Sport per Week (1 = more than 3 times a week, 4 = never)				0.337	0.288	0.738
				(0.673)	(0.681)	(0.679)
Healthy Food:much heed				−1.099	−1.126	−1.342
				(2.215)	(2.255)	(2.274)
Healthy Food:little heed				−0.416	−0.292	−1.593
				(2.216)	(2.243)	(2.280)
Healthy Food:no heed				3.100	3.707	0.731
				(3.544)	(3.676)	(3.631)
Academic Degree Parents (1 = Yes)					0.141	0.049
					(1.141)	(1.146)
Marital St. Parents:married					5.097	4.811
					(3.043)	(3.012)
Marital St. Parents:divorced					3.926	3.897
					(3.220)	(3.201)

—Continue in Next Column—

—Continued—

	(1)	(2)	(3)	(4)	(5)	(6)
Age of first exposure to second tongue					0.177	0.101
					(0.162)	(0.162)
Monthly available:between € 150 and <€ 200					−0.518	−1.793
					(1.747)	(1.760)
Monthly available:between € 200 and <€ 250					−0.674	−1.820
					(1.990)	(1.982)
Monthly available:between € 250 and <€ 300					0.608	0.172
					(2.201)	(2.189)
Monthly available:between € 300 and <€ 350					5.789	4.218
					(2.402)	(2.440)
Monthly available:between € 350 and <€ 400					1.493	1.309
					(2.493)	(2.442)
Monthly available:between € 400 and <€ 450					−3.563	−4.509
					(2.753)	(2.696)
Monthly available:between € 450 and <€ 500					1.924	−0.279
					(3.130)	(3.170)
Monthly available:between € 500 and <€ 550					−3.052	−2.260
					(2.887)	(2.733)
Monthly available:between € 550 and <€ 600					6.636	6.520
					(4.021)	(3.889)
Monthly available:between € 600 and <€ 650					−3.051	−5.165
					(3.291)	(3.575)
Monthly available: >€ 650					2.302	1.393
					(2.515)	(2.481)
Age						−0.072
						(0.154)
Gender:female						−6.571
						(1.348)
Gender: NA						−14.465
						(3.450)
Most recent math grade						−1.061
						(0.649)
Marital status:married						−4.216
						(3.152)
Marital Status:divorced						−0.837
						(6.449)
Lingual-focused Studies:True						1.370
						(1.392)
Pet:True						−0.486
						(1.263)
PF|TR		2.646	2.611	3.282	3.472	3.335
		(2.243)	(2.248)	(2.222)	(2.231)	(2.204)
PF|Present-Pref			0.016	−0.057	−0.106	0.091
			(0.996)	(0.988)	(1.001)	(0.990)
Constant	37.070***	38.284***	38.646***	29.656***	23.338***	34.440***
	(0.806)	(1.189)	(1.512)	(5.394)	(6.293)	(7.097)
Obs.: 1,137						

Note: OLS estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.

b Number of sentences in the present tense ($\{0 \dots 5\}$) selected during paragraph construction task.

c The baseline category for the variables “beer", “wine", “spirit", and “mixed-drinks" is often.

d The baseline category for monthly available is "<€ 150".

e The baseline category for “gender" is “male".

f The values for the variable “math grade" spans the interval [1,5] with “1" being the best.

A.3 Belief Vignettes

Immediacy Vignettes

Table A.8: Vignette Wording – Immediacy

PF	FF	English
	(a)	
Nachdem dort die Arbeiten bald **fertiggestellt sind**, **gibt** die Stadtverwaltung in Halle an der Saale die StraSSe am Steintor wieder frei und **hebt** dann auch die Sperrung der Dessauer StraSSe wieder **auf**. Wann endet die Sperrung der Dessauer StraSSe?	Nachdem dort die Arbeiten bald fertiggestellt sein werden, **wird die Stadtverwaltung** in Halle an der Saale die StraSSe am Steintor wieder **freigeben** und **wird** dann auch die Sperrung der Dessauer StraSSe wieder **aufheben**. Wann **wird** die Sperrung der Dessauer StraSSe **enden**?	After the work there will soon be completed, the city administration in Halle an der Saale will reopen the street at the Steintor and will then also lift the closure of Dessauer StraSSe. When will the closure of Dessauer StraSSe end?
	(b)	
Im Landkreis Mansfeld-Südharz **beginnt** bald der Ausbau von schnellem Internet. Der Kreistag **berät** dazu in der nächsten auSSerordentlichen Sitzung, der Startschuss für die Anschlussarbeiten **erfolgt** dann umgehend. Wann **folgt** der Startschuss für die Anschlussarbeiten in Mansfeld-Südharz?	Im Landkreis Mansfeld-Südharz **wird** bald der Ausbau von schnellem Internet **beginnen**. Der Kreistag **wird** dazu in der nächsten auSSerordentlichen Sitzung **beraten**, der Startschuss für die Anschlussarbeiten **wird** dann umgehend **folgen**. Wann **wird** der Startschuss für die Anschlussarbeiten in Mansfeld-Südharz **folgen**?	The expansion of high-speed Internet will soon begin in the Mansfeld-Südharz district. The district council will discuss this at its next extraordinary meeting, and the starting signal for the connection work will then follow immediately. When will the starting signal for the connection work in Mansfeld-Südharz follow?
	(c)	
Laborleiter Dr. Holger G. des BonnEconLabs erwartet, dass bald alle Teilnehmer des Labors die Stifte **zurückbringen**, die sie bei Experimenten versehentlich mitgenommen haben. Er **macht** demnächst einen Aushang am schwarzen Brett des Labors. Wann **macht** Laborleiter Dr. Holger G. einen Aushang?	Laborleiter Dr. Holger G. des BonnEconLabs erwartet, dass bald alle Teilnehmer die Stifte **zurückbringen werden**, die sie bei Experimenten versehentlich mitgenommen haben. Er **wird** demnächst einen Aushang am schwarzen Brett des Labors **machen**. Wann **wird** Laborleiter Dr. Holger G. einen Aushang **machen**?	Lab manager Dr. Holger G. of BonnEconLab expects that soon all participants will return the pens they accidentally took during experiments. He will soon make a notice on the lab's bulletin board. When will lab manager Dr. Holger G. make a notice?
	(d)	
Aufgrund des extremen Kurswachstums berichtete das Handelsblatt kürzlich über Bitcoin. Experten prognostizieren, dass ein Bitcoin innerhalb des nächsten halben Jahres einen Wert von 1100 Euro **übersteigt**. Wann **übersteigt** ein Bitcoin den Wert von 1047 Euro?	Aufgrund des extremen Kurswachstums berichtete das Handelsblatt kürzlich über Bitcoin. Experten prognostizieren, dass ein Bitcoin innerhalb des nächsten halben Jahres einen Wert von 1100 Euro **übersteigen wird**. Wann **wird** ein Bitcoin den Wert von 1047 Euro **übersteigen**?	Due to the extreme price growth, Handelsblatt recently reported on Bitcoin. Experts predict that one Bitcoin will exceed a value of 1100 euros within the next six months. When will one Bitcoin exceed the value of 1047 euros?

Notes: Wording of Immediacy Vignettes for the present-tense future reference (PF) and future-tense future reference (FF) framing. Differences betweens frames in **bold** font.

Robustness checks support these findings. We estimate a set of ordered logit models (OLM), adding a number of controls. The effect of the framing is captured by the variable PF Framing, which is 1 if the subject observed the PF Framing, and zero otherwise. For

Table A.9: Ordered Logit Estimations for Immediacy Vignettes (shortened)

	(a)	(b)	(c)	(d)
PF	0.713***	0.415*	−0.254	−0.113
	(0.208)	(0.205)	(0.220)	(0.209)
IL[a]	0.022	0.332*	0.308	−0.323*
	(0.148)	(0.150)	(0.163)	(0.155)
Present-Pref[b]	0.056	−0.006	−0.003	0.010
	(0.068)	(0.068)	(0.074)	(0.071)
PF|IL[a]	0.342	0.021	0.362	0.084
	(0.211)	(0.212)	(0.226)	(0.213)
PF|Present-Pref[b]	0.003	−0.046	0.033	0.059
	(0.097)	(0.096)	(0.104)	(0.098)
Add. Controls[c]:	Yes	Yes	Yes	Yes
		Obs.: 1,137		

Note: Ordered Logit estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.

b Number of sentences in the present tense ({0...5}) selected during paragraph construction task.

c Additional controls include: gender, alcohol consumption (beer/wine/spirits, smoking, sports per week, marital status, avalable monthly income, language-related studies, pet owner).

an overview, see Table A.9, which shows a reduced set of control variables. We depict the controls for the order in which time and Likelihood Vignettes were shown to the subjects. The elicited individual preferences for present tense, as measured by the paragraph-construction task, are also shown. Both variables are included as interactions with the respective framing as observed by each subject. Additionally, we include the responses to survey questions eliciting risk aversion, taken from the SOEP (Wagner et al., 2007). These are included to capture linkages between time and risk (Anderhub et al., 2001; Andersen et al., 2008). Finally, a standard set of socio-economic and sociodemographic controls is included. Table A.9 presents the estimation results. When including additional controls, the results remain in line with the results of the non-parametric tests. While we find some evidence for an effect of grammatical framing on time, this effect seems to be spurious, highly context-dependent, and easily disrupted.

Likelihood Vignettes

In order to test these findings on their robustness, we estimated six OLS models for each vignette, increasing the number of controls in each model subsequently, as had been done for the OLM models. The truncated results can be seen in Table A.12. In the baseline model, the results from the boxplots as well as the U-tests are validated. Similar to the

Table A.10: Ordered Logit Estimations for Immediacy Vignettes

	(a)	(b)	(c)	(d)
PF	0.713***	0.415*	−0.254	−0.113
	(0.208)	(0.205)	(0.220)	(0.209)
IL[a]	0.022	0.332*	0.308	−0.323*
	(0.148)	(0.150)	(0.163)	(0.155)
Present-Pref[b]	0.056	−0.006	−0.003	0.010
	(0.068)	(0.068)	(0.074)	(0.071)
PF|IL[a]	0.342	0.021	0.362	0.084
	(0.211)	(0.212)	(0.226)	(0.213)
PF|Present-Pref[b]	0.003	−0.046	0.033	0.059
	(0.097)	(0.096)	(0.104)	(0.098)
Self Estimation Risk (0 = fully risk averse, 11 = risk loving)	0.001	−0.018	−0.020	−0.035
	(0.029)	(0.029)	(0.031)	(0.029)
Beer:occasionally	−0.189	−0.373*	−0.263	0.034
	(0.172)	(0.171)	(0.185)	(0.174)
Beer:seldomly	0.194	−0.123	0.026	0.065
	(0.193)	(0.193)	(0.210)	(0.196)
Beer:never	0.130	−0.050	−0.087	−0.176
	(0.230)	(0.228)	(0.244)	(0.235)
Wine:occasionally	0.187	0.081	−0.099	−0.120
	(0.225)	(0.217)	(0.232)	(0.223)
Wine:seldomly	0.365	−0.100	0.057	−0.319
	(0.231)	(0.225)	(0.240)	(0.232)
Wine:never	0.344	−0.103	0.190	−0.144
	(0.254)	(0.248)	(0.266)	(0.255)
Spirit:occasionally	−0.052	−0.099	−0.663	0.395
	(0.353)	(0.338)	(0.380)	(0.344)
Spirit:seldomly	−0.405	−0.088	−0.438	0.330
	(0.355)	(0.337)	(0.382)	(0.346)
Spirit:never	−0.033	−0.001	−0.443	0.462
	(0.381)	(0.368)	(0.412)	(0.375)
Mixed-Drinks:occasionally	0.417	0.281	0.097	−0.526
	(0.400)	(0.406)	(0.448)	(0.392)
Mixed-Drinks:seldomly	0.570	0.347	0.243	−0.263
	(0.400)	(0.404)	(0.446)	(0.389)
Mixed-Drinks:never	0.697	0.216	0.064	−0.380
	(0.418)	(0.422)	(0.465)	(0.410)
Smoker (1 = Yes)	0.154	0.236	0.156	−0.112
	(0.189)	(0.195)	(0.202)	(0.195)
Sport per Week (1 = more than 3 times a week, 4 = never)	0.069	−0.111	−0.016	−0.038
	(0.061)	(0.062)	(0.066)	(0.063)
Healthy Food:much heed	0.067	−0.199	−0.169	0.108
	(0.200)	(0.194)	(0.208)	(0.198)
Healthy Food:little heed	0.107	−0.100	−0.071	0.249
	(0.201)	(0.195)	(0.210)	(0.198)
Healthy Food:no heed	−0.864*	−0.258	−0.373	0.265
	(0.366)	(0.368)	(0.398)	(0.376)
Academic Degree Parents (1 = Yes)	0.039	0.158	0.144	−0.105
	(0.109)	(0.110)	(0.117)	(0.111)
Marital St. Parents:married	0.488	−0.341	−0.467	0.164
	(0.291)	(0.299)	(0.329)	(0.294)
Marital St. Parents:divorced	0.414	−0.171	−0.355	0.202
	(0.307)	(0.315)	(0.346)	(0.311)
Age of first exposure to second tongue	0.002	−0.0003	−0.032	−0.019
	(0.016)	(0.016)	(0.017)	(0.017)
Monthly available:between € 150 and < € 200	0.343*	−0.008	−0.109	0.108
	(0.162)	(0.165)	(0.175)	(0.168)
Monthly available:between € 200 and <€ 250	0.259	0.166	0.142	−0.055
	(0.192)	(0.190)	(0.205)	(0.195)
Monthly available:between € 250 and <€ 300	−0.118	0.070	0.147	−0.246
	(0.211)	(0.208)	(0.230)	(0.214)
Monthly available:between € 300 and <€ 350	0.078	0.041	−0.166	0.143
	(0.224)	(0.225)	(0.241)	(0.221)
Monthly available:between € 350 and <€ 400	0.412	0.265	−0.032	−0.187
	(0.247)	(0.234)	(0.252)	(0.240)
Monthly available:between € 400 and <€ 450	0.232	−0.403	−0.211	0.053
	(0.248)	(0.261)	(0.259)	(0.249)
Monthly available:between € 450 and <€ 500	0.217	0.071	−0.314	0.068
	(0.308)	(0.312)	(0.329)	(0.332)
Monthly available:between € 500 and <€ 550	−0.315	−0.543	0.187	0.429
	(0.336)	(0.336)	(0.369)	(0.337)
Monthly available:between € 550 and <€ 600	0.749*	−0.106	−0.067	0.542
	(0.376)	(0.377)	(0.406)	(0.378)
Monthly available:between € 600 and <€ 650	0.262	0.251	0.532	−0.017
	(0.454)	(0.446)	(0.526)	(0.420)
Monthly available: > € 650	0.058	−0.221	−0.369	−0.020
	(0.240)	(0.243)	(0.249)	(0.236)
Age	0.024	−0.037*	0.006	0.002
	(0.013)	(0.014)	(0.014)	(0.014)
Gender:female	−0.073	−0.122	−0.612***	−0.397**
	(0.130)	(0.131)	(0.141)	(0.132)
Gender: NA	0.479	−0.152	0.023	0.049
	(0.417)	(0.427)	(0.446)	(0.410)
Most recent math grade	0.103	0.131*	0.029	−0.088
	(0.061)	(0.061)	(0.065)	(0.062)
Marital status:married	−0.659*	0.199	−0.196	−0.168
	(0.283)	(0.300)	(0.305)	(0.306)
Marital Status:divorced	0.703	2.300**	0.532	0.023
	(0.663)	(0.700)	(0.713)	(0.666)
Lingual-focused Studies:True	−0.352**	−0.198	0.147	−0.058
	(0.128)	(0.127)	(0.136)	(0.128)
Pet:True	0.102	0.050	0.105	−0.091
	(0.118)	(0.120)	(0.126)	(0.120)
less than a week|in 1 to 2 weeks	−4.890***	−2.477***	0.774	−3.681***
	(0.705)	(0.718)	(0.747)	(0.734)
in 1 to 2 weeks|in 2 to 3 weeks	−4.052***	−1.339	2.524***	−2.762***
	(0.700)	(0.708)	(0.750)	(0.702)
in 2 to 3 weeks|in 3 to 4 weeks	−3.573***	−0.656	3.380***	−2.088**
	(0.698)	(0.707)	(0.754)	(0.691)
in 3 to 4 weeks|in 4 to 5 weeks	−3.135***	−0.019	3.856***	−1.396*
	(0.696)	(0.706)	(0.757)	(0.686)
in 4 to 5 weeks|in 5 to 6 weeks	−2.835***	0.478	3.994***	−1.021
	(0.695)	(0.706)	(0.758)	(0.684)
in 5 to 6 weeks|later than 1.5 months	−2.321***	0.989	4.237***	−0.429
	(0.694)	(0.706)	(0.760)	(0.683)
later than 1.5 months|later than 2 months	−1.900**	1.443*	4.392***	0.108
	(0.693)	(0.706)	(0.762)	(0.682)
later than 2 months|later than 3 months	−1.415*	1.913**	4.523***	1.080
	(0.692)	(0.707)	(0.764)	(0.683)
later than 3 months|later than 4 months	−0.972	2.465***	4.620***	2.046**
	(0.692)	(0.709)	(0.765)	(0.686)
later than 4 months|later than 6 months	−0.562	2.927***	4.699***	3.837***
	(0.693)	(0.711)	(0.767)	(0.696)
Obs.:	1,137			

Note: Ordered Logit estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.

b Number of sentences in the present tense ({0…5}) selected during paragraph construction task.

c The baseline category for the variables “beer", “wine", “spirit", and “mixed-drinks" is often.

d The baseline category for monthly available is "<€ 150".

e The baseline category for “gender" is “male".

f The values for the variable “math grade" spans the interval [1,5] with “1" being the best.

Table A.11: Vignette Wording – Likelihood

PF	FF	English
	(e)	
Laut McKinsey **geht** in der westeuropäischen Versicherungsbranche in den nächsten zehn Jahren jeder vierte Arbeitsplatz **verloren**. Besonders davon betroffen **ist** der Bereich der Schadensabwicklung, wo jeder dritte Arbeitsplatz **verloren geht**. **Bestätigt** sich die Prognose für den Bereich der Schadensabwicklung in den nächsten zehn Jahren?	Laut McKinsey **wird** in der westeuropäischen Versicherungsbranche in den nächsten zehn Jahren jeder vierte Arbeitsplatz **verloren** gehen. Besonders davon betroffen wird der Bereich der Schadensabwicklung sein, wo jeder dritte Arbeitsplatz verloren gehen wird. **Wird** sich die Prognose für den Bereich der Schadensabwicklung in den nächsten zehn Jahren **bestätigen**?	According to McKinsey, one in four jobs will be lost in the Western European insurance industry over the next ten years. This will particularly affect the area of claims processing, where one in three jobs will be lost. Will the forecast for the claims handling sector be confirmed in the next ten years?
	(f)	
Der Chopin-Flughafen in Warschau verzeichnete ein Wachstum des Passagieraufkommens um 15% und beförderte im Kalenderjahr 2016 12,8 Millionen Passagiere. Laut Betreibergesellschaft **sind** die Zuwächse für 2017 stabil. Auch der Frachtverkehr **nimmt** dann in vergleichbarem Umfang zu. **Fertigt** der Flughafen 2017 bis zum Jahresende mindestens 14,3 Millionen Passagiere **ab**?	Der Chopin-Flughafen in Warschau verzeichnete ein Wachstum des Passagieraufkommens um 15% und beförderte im Kalenderjahr 2016 12,8 Millionen Passagiere. Laut Betreibergesellschaft **werden** die Zuwächse für 2017 stabil **sein**. Auch der Frachtverkehr **wird** dann in vergleichbarem Umfang zunehmen. **Wird** der Flughafen 2017 bis zum Jahresende mindestens 14,3 Millionen Passagiere **abfertigen**?	Chopin Airport in Warsaw recorded 15% growth in passenger traffic, carrying 12.8 million passengers in calendar year 2016. According to the operating company, the increases for 2017 will be stable. Cargo traffic will then also increase at a comparable rate. Will the airport handle at least 14.3 million passengers by the end of the year in 2017?
	(g)	
Der Bundesrechnungshof **veröffentlicht** kommende Woche einen Bericht darüber, dass alle öffentlichen Körperschaften im nächsten Jahr 5.500 neue Stellen **benötigen**. Der Bericht **stellt** dar, dass sich dennoch die Arbeit pro Kopf im öffentlichen Dienst **erhöht** und die Zahl der Krankmeldungen dadurch **ansteigt**. **Erhöht** sich die Arbeitslast pro Kopf im öffentlichen Dienst im kommenden Jahr?	Der Bundesrechnungshof **wird** kommende Woche einen Bericht darüber **veröffentlichen**, dass alle öffentlichen Körperschaften im nächsten Jahr 5.500 neue Stellen **benötigen werden**. Der Bericht **wird darstellen**, dass sich dennoch die Arbeit pro Kopf im öffentlichen Dienst **erhöhen wird** und die Zahl der Krankmeldungen dadurch **ansteigen wird**. **Wird** sich die Arbeitslast pro Kopf im öffentlichen Dienst im kommenden Jahr **erhöhen**?	The federal General Accounting Office will release a report next week showing that all public entities will need 5,500 new positions next year. The report will outline that despite this, the workload per capita in the public sector will increase and sick leave will rise as a result. Will the workload per capita in the public sector increase in the coming year?
	(h)	
In Buxtehude **beginnt** der Bau eines neuen Fahrradweges im Stadtpark. Der Förderverein Stadtpark Buxtehude e.V. **berät** bei seiner nächsten Vollversammlung über die zusätzliche Beschilderung. Sie **beraten** dabei dann ebenfalls über das Aufstellen zusätzlicher Hundekot-Beutelspendern. **Finanziert** der Förderverein das Aufstellen zusätzlicher Hundekot-Beutelspender im Buxtehuder Stadtpark?	In Buxtehude **wird** der Bau eines neuen Fahrradweges im Stadtpark **beginnen**. Der Förderverein Stadtpark Buxtehude e.V. **wird** bei seiner nächsten Vollversammlung über die zusätzliche Beschilderung **beraten**. Sie **werden** dann ebenfalls über das Aufstellen zusätzlicher Hundekot-Beutelspendern **beraten**. **Wird** der Förderverein das Aufstellen zusätzlicher Hundekot-Beutelspender im Buxtehuder Stadtpark **finanzieren**?	In Buxtehude, the construction of a new bicycle path in the city park will begin. The support association "Stadtpark Buxtehude e.V." will discuss the additional signage at its next plenary meeting. They will then also discuss the placement of additional dog waste bag dispensers. Will the support association finance the installation of additional dog waste bag dispensers in Buxtehude City Park?

Notes: Wording of Likelihood Vignettes for the present-tense future reference (PF) and future tense future reference (FF) framing. Differences betweens frames in **bold** font.

vignette "Bitcoin" in the prior section, the vignette "Buxtehude", which again brings with it a relatively strong outside predisposition, shows no significant framing effect for all estimated models. This is seen as support for the assumption that any grammatical framing effect, irrespectively of the domain targeted, is susceptible to preconceived opinions or outside predispositions and additional information.

As before, the framing effects prove unstable when including additional control variables. This can especially be seen for the case of the "BonnEconLab" vignette, for which the framing effect loses any significance when including the additional controls.

Table A.12: OLS Estimations for Likelihood Vignettes (shortened)

	(e)	(f)	(g)	(h)
PF	8.044**	1.100	2.531	−1.043
	(2.704)	(3.033)	(2.990)	(2.952)
IL[a]	4.847*	−3.133*	4.844	−1.327
	(1.908)	(2.164)	(2.183)	(2.217)
Present-Pref[b]	0.717	0.629	0.401	−0.129
	(0.868)	(1.034)	(1.034)	(0.964)
PF|IL[a]	−3.123	3.653	−3.387	2.963
	(2.766)	(3.167)	(3.054)	(3.098)
PF|Present-Pref[b]	−1.809	−0.160	1.389	−0.626
	(1.311)	(1.503)	(1.406)	(1.365)
Add. Controls[c]:	Yes	Yes	Yes	Yes
		Obs.: 1,137		

Note: OLS. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.

b Number of sentences in the present tense ({0...5}) selected during paragraph construction task.

c Additional controls include: gender, alcohol consumption (beer/wine/spirits, smoking, sports per week, marital status, avalable monthly income, language-related studies, pet owner). are available from the authors upon request.

Consequently, we do not find a stable effect on risk due to grammatical framing either.

A.4 Experimental Material

Due to the fact that Amazon mTurk only became available after we conducted our experiment, we opted for using a lab subject pool in an online experiment. This brings with it the advantage of having a reliable subject pool used to economic decision-making and different experimental setups. Accordingly, we also complied with the strict lab rules in all matters such as anonymity and average payoff requirements. The decision was

Table A.13: OLS Estimations for Likelihood Vignettes

	(e)	(f)	(g)	(h)
PF	8.044**	1.100	2.531	−1.043
	(2.704)	(3.033)	(2.990)	(2.952)
IL[a]	4.847*	−3.133*	4.844	−1.327
	(1.908)	(2.164)	(2.183)	(2.217)
Present-Pref[b]	0.717	0.629	0.401	−0.129
	(0.868)	(1.034)	(1.034)	(0.964)
Self Estimation Risk (0 = fully risk averse, 11 = risk loving)	0.605	−0.113	−0.510	0.435
	(0.389)	(0.441)	(0.441)	(0.422)
Beer:occasionally	−3.012	−2.961**	−6.700	−2.903
	(2.236)	(2.550)	(2.403)	(2.434)
Beer:seldomly	−4.209	−3.015	−5.354	−0.832
	(2.573)	(2.860)	(2.850)	(2.747)
Beer:never	−2.313	1.452	−3.042	−3.064
	(3.041)	(3.324)	(3.151)	(3.397)
Wine:occasionally	5.297*	−0.834	1.888	−0.018
	(2.647)	(3.076)	(2.854)	(2.934)
Wine:seldomly	5.615*	−1.099	2.176	−2.269
	(2.836)	(3.222)	(2.925)	(3.055)
Wine:never	4.274	−1.683	−0.859	−3.066
	(3.156)	(3.647)	(3.279)	(3.466)
Spirit:occasionally	−5.127	6.426	6.837	−7.197
	(4.435)	(5.557)	(4.936)	(4.507)
Spirit:seldomly	−3.809	6.650	7.833	−4.086
	(4.451)	(5.646)	(4.970)	(4.527)
Spirit:never	−3.506	3.841	6.189	−7.580
	(4.900)	(6.020)	(5.371)	(5.061)
Mixed-Drinks:occasionally	−0.135	−0.876	−5.462	17.879***
	(5.199)	(5.460)	(4.894)	(4.955)
Mixed-Drinks:seldomly	3.644	−2.710	−3.314	15.117**
	(5.140)	(5.443)	(4.910)	(4.930)
Mixed-Drinks:never	2.958	0.187	−3.020	14.588**
	(5.428)	(5.630)	(5.211)	(5.239)
Smoker (1 = Yes)	−4.080	1.723*	5.977	−7.158*
	(2.472)	(2.607)	(2.700)	(3.093)
Sport per Week (1 = more than 3 times a week, 4 = never)	0.056	0.315	1.085	0.386
	(0.809)	(0.952)	(0.876)	(0.860)
Healthy Food:much heed	−3.296	0.653	1.747	0.674
	(2.531)	(3.018)	(2.926)	(2.941)
Healthy Food:little heed	−0.223	0.624	−0.417	0.117
	(2.555)	(3.088)	(2.927)	(2.968)
Healthy Food:no heed	−7.012	−0.946	0.211	0.241
	(4.465)	(5.289)	(5.405)	(5.168)
Academic Degree Parents (1 = Yes)	1.467	2.736**	4.589	−1.027
	(1.408)	(1.641)	(1.601)	(1.596)
Marital St. Parents:married	1.162	−3.048	−0.778	−6.055
	(4.260)	(4.167)	(4.512)	(4.328)
Marital St. Parents:divorced	−0.650	−1.753	0.898	−6.369
	(4.485)	(4.396)	(4.762)	(4.551)
Age of first exposure to second tongue	0.069	0.173	−0.079	−0.086
	(0.205)	(0.242)	(0.240)	(0.232)
Monthly available:between € 150 and < € 200	1.470	−0.539	−1.333	2.264
	(2.211)	(2.431)	(2.367)	(2.447)
Monthly available:between € 200 and <€ 250	4.055	0.392	−4.166	2.231
	(2.512)	(3.018)	(2.682)	(2.740)
Monthly available:between € 250 and <€ 300	0.698	0.933	−2.105	1.519
	(2.866)	(3.162)	(2.934)	(3.077)
Monthly available:between € 300 and <€ 350	1.892	−0.698*	−6.777	1.761
	(2.772)	(3.446)	(3.387)	(3.095)
Monthly available:between € 350 and <€ 400	−1.927	−3.157	−6.531	−1.005
	(3.079)	(3.850)	(3.628)	(3.655)
Monthly available:between € 400 and <€ 450	7.596*	1.690	0.455	3.248
	(3.129)	(3.388)	(3.226)	(3.717)
Monthly available:between € 450 and <€ 500	3.748	−2.646	0.294	0.663
	(3.836)	(4.857)	(4.669)	(4.519)
Monthly available:between € 500 and <€ 550	6.459	5.250	0.473	−0.283
	(4.965)	(4.773)	(4.465)	(3.996)
Monthly available:between € 550 and <€ 600	1.923	−3.465	−9.727	4.133
	(4.178)	(6.502)	(5.988)	(5.338)
Monthly available:between € 600 and <€ 650	12.026*	6.344	4.681	0.048
	(5.417)	(5.505)	(5.730)	(5.752)
Monthly available: > € 650	−1.519	−2.033	−4.090	1.577
	(3.001)	(3.627)	(3.388)	(3.298)
Age	0.142	0.024*	0.447	−0.195
	(0.183)	(0.197)	(0.188)	(0.176)
Gender:female	−1.515	−1.632	−2.023	0.701
	(1.719)	(2.042)	(1.877)	(1.911)
Gender: NA	−3.317	−3.242	−2.785	−0.460
	(5.025)	(6.741)	(6.600)	(6.359)
Most recent math grade	0.655	−0.984	−1.139	−0.269
	(0.802)	(0.863)	(0.884)	(0.865)
Marital status:married	−0.499	−3.156	−4.616	2.348
	(3.602)	(4.106)	(4.501)	(4.048)
Marital Status:divorced	−9.082	−2.361	−3.114	15.300*
	(8.744)	(8.174)	(8.827)	(6.215)
Lingual-focused Studies:True	1.012	−2.591	1.962	0.583
	(1.593)	(1.890)	(1.824)	(1.788)
Pet:True	−1.720	−0.323	−1.136	−0.106
	(1.516)	(1.796)	(1.770)	(1.755)
PF|IL[a]	−3.123	3.653	−3.387	2.963
	(2.766)	(3.167)	(3.054)	(3.098)
PF|Present-Pref[b]	−1.809	−0.160	1.389	−0.626
Obs.: 1,137				

—Continue in Next Column—

—Continued—

Note: OLS estimation results. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

a Interaction effect between the block-wise order of Immediacy and Likelihood Vignettes and PF framing. Order Immediacy → Risk = Likelihood = 1.
b Number of sentences in the present tense ({0…5}) selected during paragraph construction task.
c The baseline category for the variables “beer", “wine", “spirit", and “mixed-drinks" is often.
d The baseline category for monthly available is "<€ 150".
e The baseline category for “gender" is “male".
f The values for the variable “math grade" spans the interval [1,5] with “1" being the best.

made to structure the payments as a lottery which means only those subjects drawn in the lottery were payed according to their choices made during the experiment. This was made clear to subjects before they could design to apply for the experiment. While it could potentially lead to subjects not taking the experiments seriously and result in random decisions, we compensate for that by scaling the payments accordingly. Subjects drawn in the lottery could earn between 120 Euro and 200 Euro instead of the 10 Euro - 20 Euro for usual lab experiments. This kept the average payment per subject up to usual lab standards.

Instructions

Below you find the set of instructions. As the experiment varied specific grammatical properties of the instructions written in the German language we provide translations upon request.

Table A.14: Instructions Introduction and Risk Choice Task

PF	FF	English
	Welcome	
Sie **nehmen** an einer wirtschaftswissenschaftlichen Studie **teil**. Bitte lesen Sie alle Erklärungen in dieser Studie aufmerksam und genau durch. Es ist wichtig, dass Sie alle Erklärungen genau verstehen, da Sie im Folgenden abhängig von Ihren Entscheidungen die Möglichkeit **haben**, eine beträchtliche Summe Geld zu erhalten. Wie viel Geld Sie am Ende der Studie tatsächlich **erhalten**, **wird** von Ihren Entscheidungen sowie einer Lotterie **abhängen**. Wir **verlosen** unter allen Teilnehmern mehrere Auszahlungen. Nur die Lotteriegewinner **erhalten** eine Auszahlung. Wir **weisen** Sie auf den folgenden Seiten nicht erneut auf die Lotterie hin. Damit Sie nicht aus Versehen auf den 'Weiter'-Button klicken, **erscheint** der Button jeweils mit einer kurzen zeitlichen Verzögerung.	Sie **werden** an einer wirtschaftswissenschaftlichen Studie **teilnehmen**. Bitte lesen Sie alle Erklärungen in dieser Studie aufmerksam und genau durch. Es ist wichtig, dass Sie alle Erklärungen genau verstehen, da Sie abhängig von Ihren Entscheidungen die Möglichkeit **haben werden**, eine beträchtliche Summe Geld zu erhalten. Wie viel Geld Sie am Ende der Studie tatsächlich **erhalten werden**, **wird** von Ihren Entscheidungen sowie einer Lotterie **abhängen werden**. Wir **werden** unter allen Teilnehmern mehrere Auszahlungen **verlosen**. Nur die Lotteriegewinner **werden** eine Auszahlung **erhalten**. Wir **werden** Sie auf den folgenden Seiten nicht erneut auf die Lotterie **hinweisen**. Hinweis: Damit Sie nicht aus Versehen auf den 'Weiter'-Button klicken, **wird** der Button jeweils mit einer kurzen zeitlichen Verzögerung **erscheinen**.	You will participate in an economic study. Please read all explanations in this study carefully and thoroughly. It is important that you understand all the explanations carefully, because depending on your decisions you will have the possibility to receive a considerable amount of money. How much money you will actually receive at the end of the study will depend on your decisions as well as a lottery. We will draw several payouts among all participants. Only the lottery winners will receive a payout. We will not mention the lottery again on the following pages. Note: To prevent you from accidentally clicking on the 'Continue' button, the button will appear with a short time delay in each case.
	BombGame	
In diesem Teil der Studie, "Meide die Bombe", **klicken** Sie Felder an. Für jedes Feld, das Sie **auswählen**, **erhalten** Sie 0.20 Euro. Das Spielfeld **besteht** aus 100 Feldern. Unter einem der Felder **befindet** sich eine verdeckte Bombe. Das Bombenfeld **wird** zufällig **bestimmt**, wobei alle Felder die gleiche Wahrscheinlichkeit **haben** die Bombe zu enthalten. Wenn sich die Bombe unter einem von Ihnen gewählten Feld **befindet**, **erhalten** Sie keine Auszahlung aus diesem Spiel. (0.00 Euro) Ist die Bombe nicht in den ausgewählten Feldern enthalten, **erhalten** Sie (Anzahl der ausgewählten Felder) * 0.20 Euro. Ob Sie die Bombe ausgewählt haben, **erfahren** Sie auf einer späteren Seite.	In diesem Teil der Studie, "Meide die Bombe", **werden** Sie Felder **anklicken**. Für jedes Feld, das Sie **auswählen werden**, **werden** Sie 0,20 Euro **erhalten**. Das Spielfeld **wird** aus 100 Feldern **bestehen**. Unter einem der Felder **wird** sich eine verdeckte Bombe **befinden**. Das Bombenfeld **wird zufällig bestimmt werden**, wobei alle Felder die gleiche Wahrscheinlichkeit **haben werden** die Bombe zu enthalten. Wenn sich die Bombe unter einem von Ihnen gewählten Feld befindet, **werden** Sie keine Auszahlung aus diesem Spiel **erhalten**. (0.00 Euro) Ist die Bombe nicht in den ausgewählten Feldern enthalten, **werden** Sie (Anzahl der ausgewählten Felder) * 0.20 Euro **erhalten**. Ob Sie die Bombe ausgewählt haben, **werden** Sie auf einer späteren Seite **erfahren**.	In this part of the study, "Avoid the bomb", you will select fields. For each field you clicked, you will receive 0.20 Euro. The game field will consist of 100 squares. Under one of the squares there will be a hidden bomb. The bomb field will be determined randomly, and all the fields will have the same probability of containing the bomb. If the bomb is under one of your chosen squares, you will not receive any payout from this game. (0.00 Euro) If the bomb is not in the selected fields, you will receive (number of selected fields) * 0.20 Euro. You will find out if you have selected the bomb on a later page.

Notes: Wording of introductions and experimental instructions for the Bomb Game. The table shows the original present-tense future reference (PF) and future-tense future reference (FF) framing as well as the English translation

Table A.15: Instructions Time Choice Task

PF	FF	English
	Choice List	
In diesem Teil der Studie **wählen** Sie eine Auszahlungsreihe aus den 10 angebotenen Auszahlungsreihen aus. Diese Auszahlungsreihen **sind** unabhängig von potenziellen Auszahlungen in anderen Teilen der Studie. Die Auszahlungen für jede Reihe **finden** jeweils über eine Dauer von 6 Wochen **statt**, aber **beginnen** zu unterschiedlichen Zeitpunkten. Wir **geben** zu den in der Auszahlungsreihe dargestellten Terminen jeweils eine Banküberweisung an Sie in Auftrag.	In diesem Teil der Studie **werden** Sie eine Auszahlungsreihe aus den 10 angebotenen Auszahlungsreihen **auswählen**. Diese Auszahlungsreihen **werden** unabhängig von potenziellen Auszahlungen in anderen Teilen der Studie **sein**. Die Auszahlungen für jede Reihe**werden** jeweils über eine Dauer von 6 Wochen **stattfinden**, aber **werden** zu unterschiedlichen Zeitpunkten **beginnen**. Wir **werden** zu den in der Auszahlungsreihe dargestellten Terminen jeweils eine Banküberweisung an Sie in Auftrag **geben**.	In this part of the study you will choose a payout series from the 10 offered payout series. These payout series will be independent of potential payouts in other parts of the study. The payouts for each series will take place over a period of 6 weeks, but will start at different times. We will initiate a bank transfer to you on each of the dates shown in the payout series
Kurze Übersicht:	Kurze Übersicht:	Short Overview
Auszahlungsreihe 1 **beginnt** in der kommenden Woche und Sie **erhalten** 27.95 Euro pro Woche.	Auszahlungsreihe 1 **wird** in der kommenden Woche **beginnen** und Sie **werden** 27.95 Euro pro Woche **erhalten**.	Payout series 1 will start next week and you will receive 27.95 euros per week.
Auszahlungsreihe 2 **beginnt** in zwei Wochen und Sie **erhalten** Euro 28.64 pro Woche.	Auszahlungsreihe 2 **wird** in zwei Wochen **beginnen** und Sie **werden** 28.64 Euro pro Woche **erhalten**.	Payout series 2 will start in two weeks and you will receive 28.64 euros per week.
Auszahlungsreihe 3 **beginnt** in drei Wochen und Sie **erhalten** Euro 29.26 pro Woche.	Auszahlungsreihe 3 **wird** in drei Wochen **beginnen** und Sie **werden** 29.26 Euro pro Woche **erhalten**.	Payout series 3 will start in three weeks and you will receive 29.26 euros per week.
Auszahlungsreihe 5 **beginnt** in fünf Wochen und Sie **erhalten** Euro 30.26 pro Woche.	Auszahlungsreihe 5 **wird** in fünf Wochen **beginnen** und Sie **werden** 30.26 Euro pro Woche **erhalten**.	Payout series 5 will start in five weeks and you will receive 30.26 euros per week.
Auszahlungsreihe 6 **beginnt** in sechs Wochen und Sie **erhalten** Euro 30.64 pro Woche.	Auszahlungsreihe 6 **wird** in sechs Wochen **beginnen** und Sie **werden** 30.64 Euro pro Woche **erhalten**.	Payout series 6 will start in six weeks and you will receive 30.64 euros per week.
Auszahlungsreihe 7 **beginnt** in sieben Wochen und Sie **erhalten** Euro 30.95 pro Woche.	Auszahlungsreihe 7 **wird** in sieben Wochen **beginnen** und Sie **werden** 30.95 Euro pro Woche **erhalten**.	Payout series 7 will start in seven weeks and you will receive 30.95 euros per week.
Auszahlungsreihe 8 **beginnt** in acht Wochen und Sie **erhalten** Euro 31.18 pro Woche.	Auszahlungsreihe 8 **wird** in acht Wochen **beginnen** und Sie **werden** 31.18 Euro pro Woche **erhalten**.	Payout series 8 will start in eight weeks and you will receive 31.18 euros per week.
Auszahlungsreihe 9 **beginnt** in neun Wochen und Sie **erhalten** Euro 31.33 pro Woche.	Auszahlungsreihe 9 **wird** in neun Wochen **beginnen** und Sie **werden** 31.33 Euro pro Woche **erhalten**.	Payout series 9 will start in nine weeks and you will receive 31.33 euros per week.
Auszahlungsreihe 10 **beginnt** in zehn Wochen und Sie **erhalten** Euro 31.40 pro Woche.	Auszahlungsreihe 10 **wird** in zehn Wochen **beginnen** und Sie **werden** 31.40 Euro pro Woche **erhalten**.	Payout series 10 will start in ten weeks and you will receive 31.40 euros per week.
Hinweis: Klicks auf die Buttons **zeigen** die Auszahlungsreihen im Detail **an**.	Hinweis: Klicks auf die Buttons **werden** die Auszahlungsreihen im Detail **anzeigen**.	Note: clicks on the buttons will show the payout series in detail.

Notes: Wording of experimental instructions for the Time Choice Task. The table shows the original present-tense future reference (PF) and future-tense future reference (FF) framing as well as the English translation

Table A.16: Instructions Belief Elicitation

PF	FF	English
	Likelihood	
Wie sicher sind sich Ihrer Meinung nach die anderen Teilnehmer, die diesen Text lesen im Durchschnitt, dass das in der Frage dargestellte Ereignis **eintritt**. Wenn Sie den Durchschnitt der anderen Teilnehmer genau treffen, **erhalten** Sie für diese Frage 2 Euro. Weichen Sie weniger als 5% vom Durchschnitt, ab **erhalten** Sie 1 Euro. Beachten Sie, dass es jedoch für die Beantwortung der unten gestellten Fragen kein Richtig oder Falsch gibt. Hinweis: Klicken sie auf den rot-grünen Balken um einen Prozentwert auszuwählen. Alternativ können Sie Ihre Antwort auch direkt in "Sie halten es für zu [Eingabe]% sicher." eingeben	Wie sicher sind sich Ihrer Meinung nach die anderen Teilnehmer, die diesen Text lesen im Durchschnitt, dass das in der Frage dargestellte Ereignis **eintreten wird**. Wenn Sie den Durchschnitt genau treffen, **werden** Sie für diese Frage 2 Euro **erhalten**. Weichen Sie weniger als 5% vom Durchschnitt ab, **werden** Sie 1 Euro **erhalten**. Beachten Sie, dass es jedoch für die Beantwortung der unten gestellten Fragen kein Richtig oder Falsch gibt. Hinweis: Klicken sie auf den rot-grünen Balken um einen Prozentwert auszuwählen. Alternativ können Sie Ihre Antwort auch direkt in "Sie halten es für zu [Eingabe]% sicher." eingeben	On average, how sure do you think the other participants reading this text are that the event depicted in the question will happen? If you hit the average exactly, you will receive 2 euros for this question. If you deviate less than 5% from the average, you will receive 1 euro. Please be aware that there is no right or wrong answer to the questions below. Note: Click on the red-green bar to select a percentage. Alternatively, you can enter your answer directly in "You think it is [enter]% safe."
	Immediacy	
Bitte geben Sie an, welche Option Ihrer Meinung nach die meisten Teilnehmer **wählen**, die diesen Text lesen. Wenn Sie die Option, die die meisten Teilnehmer**wählen**, richtig bestimmen, **erhalten** Sie für diese Frage zusätzlich 1 Euro. Beachten Sie, dass es für die Beantwortung der unten gestellten Fragen kein Richtig oder Falsch gibt.	Bitte geben Sie an, welche Option Ihrer Meinung nach die meisten Teilnehmer **wählen werden**, die diesen Text lesen. Wenn Sie die Option, die die meisten Teilnehmer **wählen werden**, richtig **bestimmen**, **werden** Sie für Ihre Antwort zusätzlich 1 Euro erhalten. Beachten Sie, dass es für die Beantwortung der unten gestellten Fragen kein Richtig oder Falsch gibt.	Please indicate which option you think most participants reading this text will choose. If you correctly determine the option that most participants will choose, you will receive an additional 1 euro for your answer. Please be aware that there is no right or wrong for answering the questions below.
	Paragraph Construction Task - No Framing in Instructions	
In den fünf grauen Boxen unten finden Sie jeweils 2 Sätze. Bitte wählen Sie in jeder Box aus, welcher Satz sich für Sie natürlicher anfühlt. In der blauen Box darunter wird Ihnen der Gesamttext angezeigt, der sich aus den von Ihnen gewählten Sätzen ergibt. Diser sollte sich für Sie ebenfalls möglichst natürlich anfühlen. Falls Sie Ihre Anwort ändern möchten, könne Sie dies vor dem Abschicken jederzeit tun. Auch hier gilt, es gibt kein Richtig oder Falsch. Hinweis: Sie können Ihre Auswahlen durch Anklicken der Boxentitel jederzeit noch einmal aufrufen und ändern.		In the five gray boxes below you will find 2 sentences in each one. In each box, please select which sentence feels more natural to you. In the blue box below, you'll see the overall text that results from the sentences you've chosen. This should also feel as natural to you as possible. If you want to change your answer, you can do so at any time before submitting. Again, there is no right or wrong. Note: You can revisit and change your selections at any time by clicking on the box titles.

Notes: Wording of experimental instructions for the present-tense future reference (PF) and future-tense future reference (FF) framing.

Screenshots

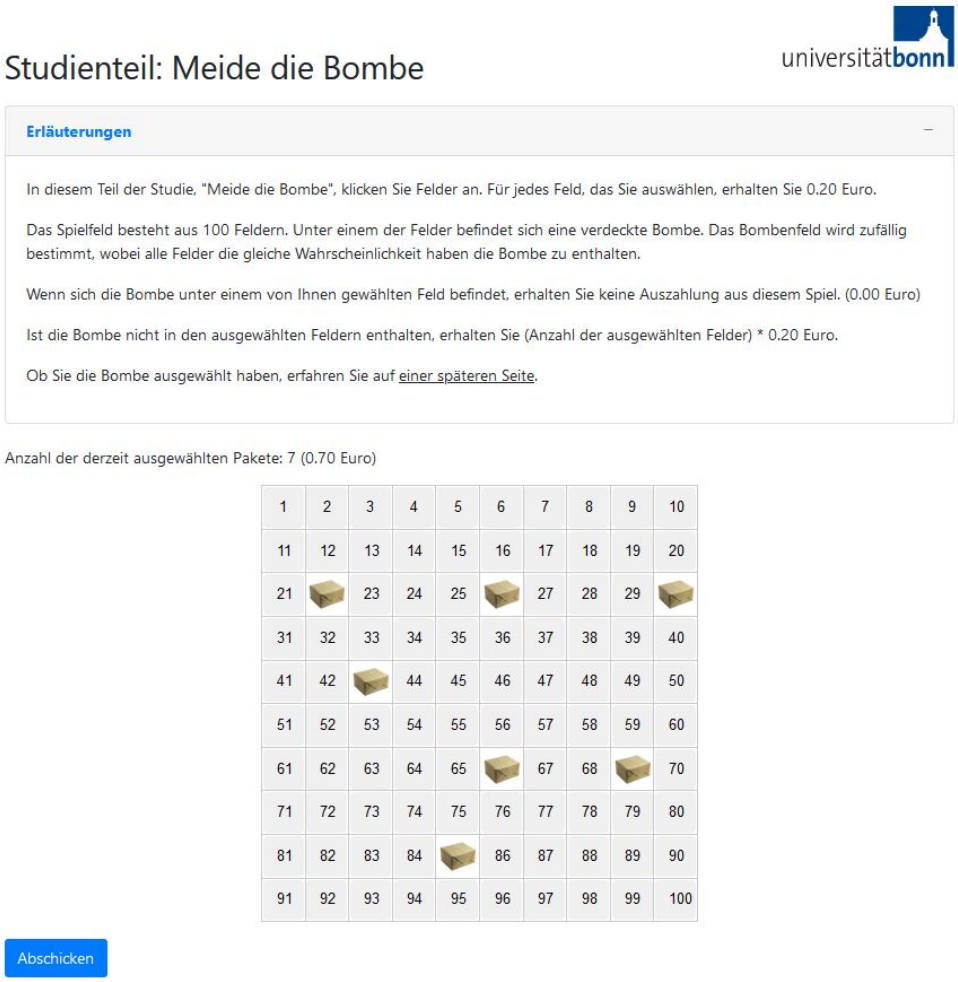


Figure A.1: Risk Preferences Elicitation Task - BombGame

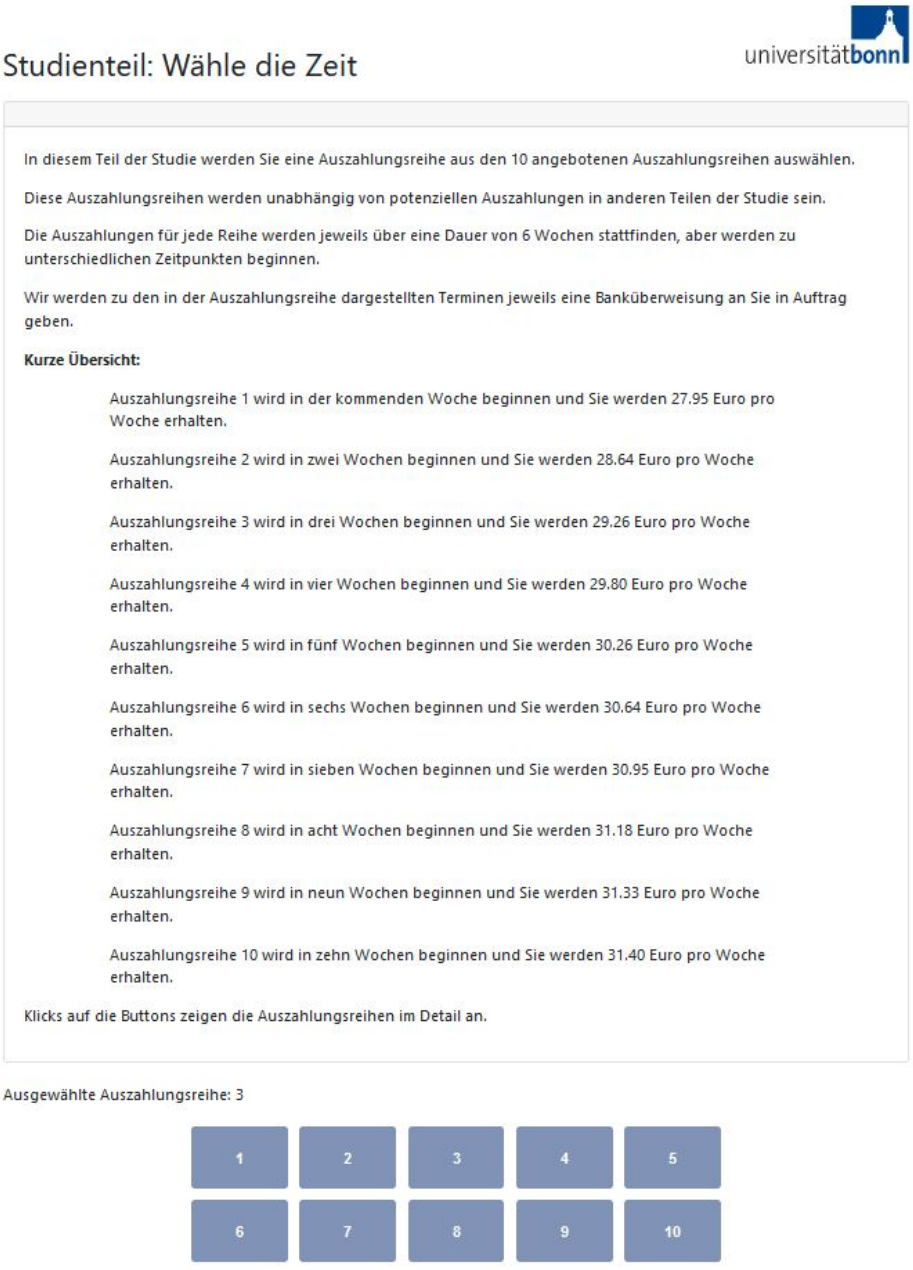


Figure A.2: Time-Preferences Elicitation Task - Choice Menu

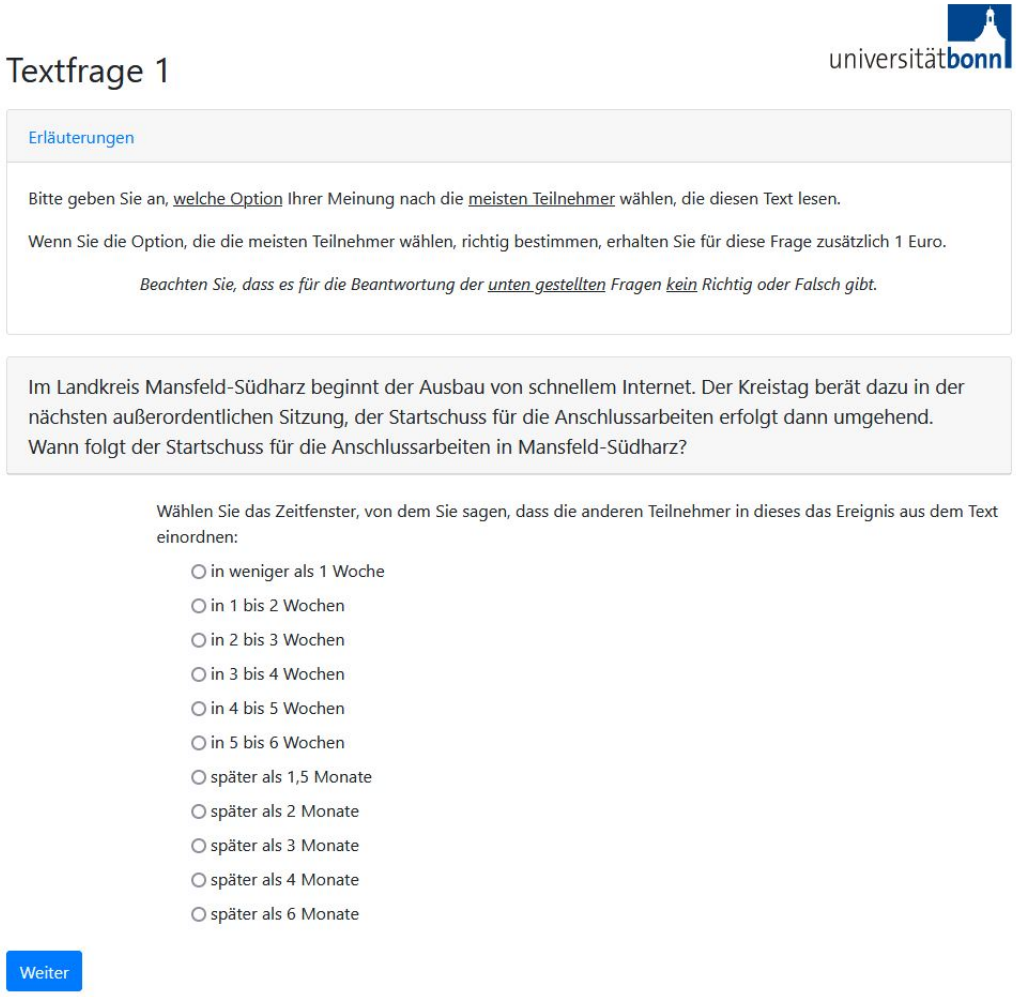


Figure A.3: Belief Elicitation Task - Immediacy Vignette

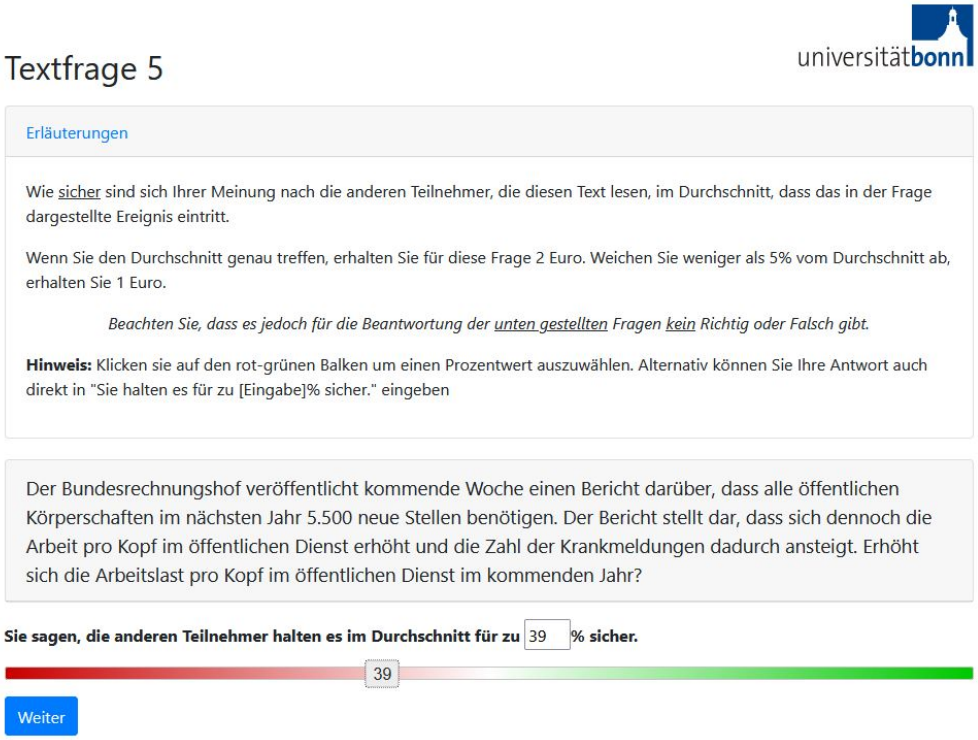


Figure A.4: Belief Elicitation Task - Likelihood Vignette

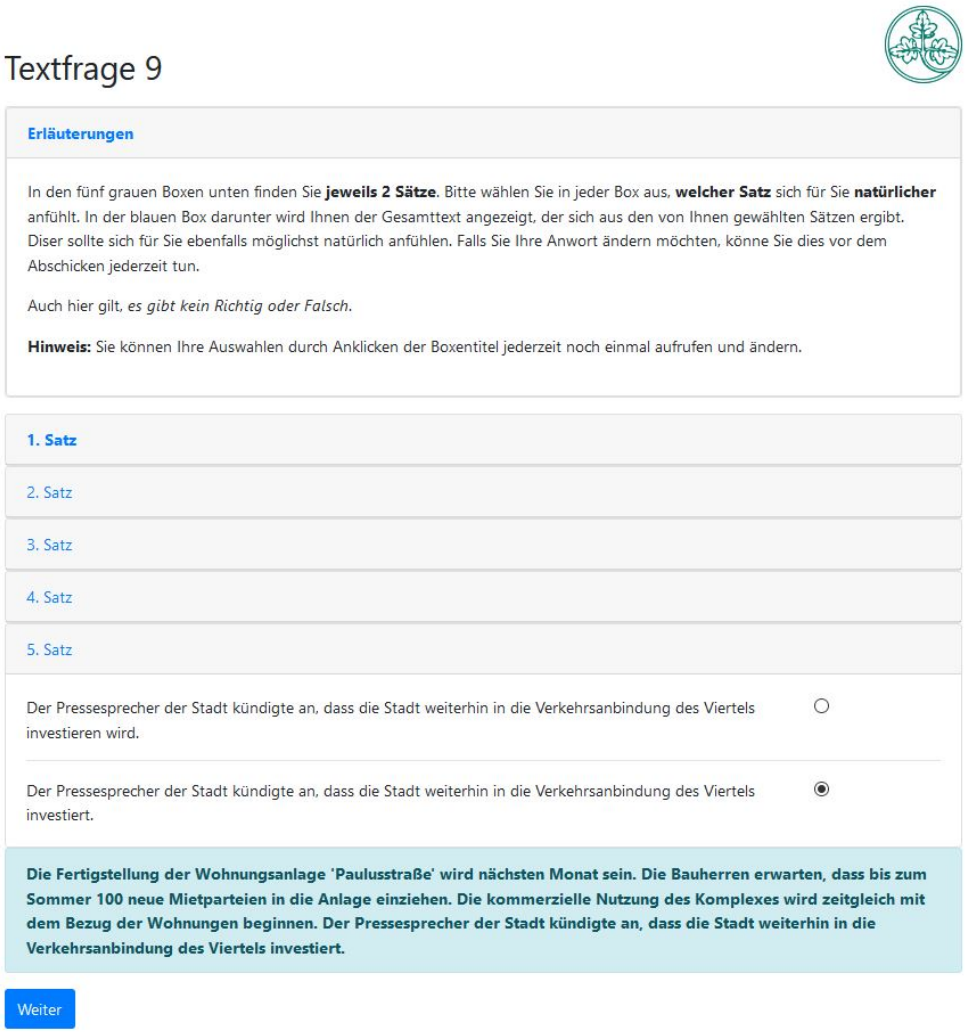


Figure A.5: Paragraph Construction Task

A.5 Survey

Table A.17: Survey Questions Targeting Risk Preferences

Question	Answer Possibilities
How do you personally rate yourself: Are you generally a risk-taking person, or do you try to avoid risk? (0: not at all willing to take risks; 10 very willing to take risks)	0-10
How often per week do you drink beer?	regularly, now and then, seldomly, never
How often per week do you drink wine?	regularly, now and then, seldomly, never
How often per week do you drink spirits?	regularly, now and then, seldomly, never
How often per week do you drink mixed alcoholic drinks?	regularly, now and then, seldomly, never
Are you a smoker currently?	yes, no
How many cigarettes do you smoke per day?	1-99
How often do you actively engage in sports, fitness or gymnastics?	three or more times per week, one or two times per week, at maximum once time every two weeks, never
To what extent do you pay attention to health-conscious nutrition?	not at all, a little bit, strongly, very strongly

Table A.18: Survey Questions Targeting Demographics and Personal Circumstances

Question	Answer Possibilities
What is your sex?	male, female, no disclosure
If applicable, please indicate your field of study or work environment. In the case of a doctorate, please indicate your doctoral field	This question does not apply to me, Business Administration, Biology, Chemistry, Computer Science, Law, Geography, History, Linguistics, Mathematics, Pharamcy, Physics, Political Sciences, Psychology, Theology, Sociology or Eductational Sciences, Economics, Other (please indicate)
Please enter the postal code of your hometown, i.e. the place where you grew up.	numerical free text
Please enter your last average grade in mathematics here. (Scale: 1.0-6.0)	1-6
How much money do you have at your free disposal each month (after deduction of rent, insurance and food)?	less than 150, between 150 and under 200, between 200 and under 250, between 250 and under 300, between 300 and under 350, between 350 and under 400, "between 400 and under 450, between 450 and under 500, between 500 and under 550, between 550 and under 600, between 600 and under 650, 650 and more
Do your parents have an academic degree (university or college)?	yes, no
Please state your marital status	unmarried, married, widowed
Please state your parent's marital status	unmarried, married, widowed
Is German your mother tongue?	yes, no
At what age did you first have intensive contact with a foreign language? (e.g.: grew up bilingually, learned through play in the kindergarden, elementary school class, etc.)	0-age
Do you have a pet?	yes, no
If yes, which kind of pet do you have?	free text

APPENDIX B

APPENDIX – CHAPTER 2

B.1 Data-Generating Process, Simulated Data, and Grid Search

For our grid search, we assumed the data-generating process of the standard public-good game, as laid out in the main section of the paper.

Simulated Type Space In their seminal paper, Fischbacher, Gächter, and Fehr (2001) argue that (in their one-shot version of this game) there are three types: free-riders, conditional cooperators, and "hump-shaped" players. In his reanalysis of Fischbacher, Gächter, Bardsley, et al. (2010), Engel (2020) further finds a small, but discernible fraction of altruists. In their reanalysis of Fischbacher, Gächter, Bardsley, et al. (2010), Engel and Rockenbach (2020) use a combination of belief and choice data to distinguish a fifth group, which they call far-sighted free-riders. In our simulations, we allow for these five types. We focus on a partner design. Groups stay together for the full duration of the game. We always allow for an individual random effect η_i and residual error $\epsilon_{it} \perp \eta_i$, which we both define to be normally distributed with mean 0 and standard deviation .3 ($\sim \mathcal{N}(0, .3)$). We thus implement the type space as defined in Table 2.1, where $c_{-i,t-1}$ is the average contribution of the remaining group members in the previous period.

We have two types that exhibit variance (between participants due to η_i, and within participants due to ϵ_{it}), but do not react to experiences: short-sighted free-riders and altruists. The contributions of these types do not have a trend either. They are random walks, albeit with diametrically opposed starting points. By contrast, the remaining three types are reactive, which may, depending on the choices of the remaining group members $c_{-i,t-1}$, lead to a trend. We have (true) conditional cooperators start in the middle of the action space. In early periods ($t < \tau = 5$), far-sighted free-riders mimic conditional cooperators, but from period τ on, they free-ride. Such participants "feed the cow" for a while, only to "start milking" it then. Finally, we simulate hump-shaped

participants such that they start rather low, at 5, and have them behave like conditional cooperators as long as the remaining group members, in the previous period, has on average not contributed more than half of the endowment. If $c_{-i,t-1} > 10$, they exhibit a perverse reaction. The more others have contributed, the less they contribute themselves.

We have groups of size $K = 4$, and we allow for $n = 5$ types. Participants choose their contributions to the public good simultaneously, which is why their order does not matter. We consider the possibility that types are present more than once in a group. Hence, we have a problem of unordered sampling with replacement. This gives us a total type space of

$$N = \binom{n+k-1}{k} = \frac{(5+4-1)!}{(5-1)!4!} = 70 \tag{B.1}$$

different group combinations. In our simulations, we include each of these 70 combinations of types N times. As three of the five types (conditional cooperators, far-sighted free-riders, hump-shaped players) are reactive, we give the classification algorithm access to the exact same experiences that participants make in this design, i.e, the mean contribution of the remaining group members in the previous period. Hence, the object of clustering is a two-dimensional time series.

- Profit is given by (2.1), with $e = 20, K = 4, t = 10$. We do, however, only use data from periods 2 - 10 for analysis, as participants have not made any experiences in the first period.
- Each of the five types is represented in equal proportions in the population from which choices are drawn.

The following elements of the data-generating process as well as of the clustering process are kept fixed for each point in the grid:

- Individual specific error $\eta \in \{0.6, 0.7, 0.8, 0.9\}$ and residual error $\sigma \in \{0.6, 0.7, 0.8, 0.9\}$.
- Sample size $n = 70 * N | N \in \{1, 2, 3, 4, 5, 6\}$
- The number of clusters $k \in \{5, 8, 10, 15, 20, 25, 30, 35, 40, 42, 44, 46, 48, 50, 52, 54\}$
- The size of the window within which dynamic time-warping is executed, running from $w \in \{1, 2, 3\}$.

The grid search thus runs over $4\eta \cdot 4\sigma \cdot 6N \cdot 3w \cdot 3\gamma \cdot 16k = 4608$ variations. The following parameters of the algorithm are varied:

- The distance used, i.e., DTW, sDTW, GAK
- The smoothing parameter $\gamma \in \{0.001, 0.01, 0.1\}$.
- The centroid function, i.e., either PAM or DBA

Each point in the dataset point in the grid is thus clustered with ten variations in the configuration of the clustering algorithm. In the simulated dataset, we have 1,120 pairs of time series for experiences and individual choices. If it were necessary to estimate 350 clusters, there would be little more than 3 participants per cluster, on average. The simulated dataset would be too small for the purpose. More disturbingly, it would be very difficult to compile a set of experimental data that is big enough for the ultimate goal of this study: to find out whether there are untheorized types.

In the interest of finding the appropriate number of clusters, and of better understanding the relationship between the degree of smoothing and the number of clusters that best organize the data, we evaluate the simulated dataset.

35 clusters for 5 types A comparison between Figure 2.1b and Figure B.1 shows how important it is to increase the number of clusters. The algorithm can still not perfectly discriminate between the types that have generated experiences and choices. This in particular holds for clusters 25—28. In these clusters, contributions and experiences are similar. They start at a differently high level and gradually decay. This pattern is generated by conditional cooperators, hump-shaped players, and (in cluster 25) also far-sighted free-riders interacting with each other. If they are in a group that, otherwise, is very cooperative, hump-shaped players and short-sighted free-riders generate a similar pattern (cluster 23). Conditional cooperators and hump-shaped players look the same if they are in a group dominated by far-sighted free-riders (cluster 13) or by short-sighted free-riders (cluster 24). Altruists and conditional cooperators are lumped together if the group quickly converges to full cooperation (cluster 9). However, even in all these clusters, while types are not perfectly separated, patterns are very cleanly characterized. The algorithm visibly does a very good job. Types are not distinguished because different reaction functions generate choice patterns that are very similar, provided a participant makes the experiences defined in the respective right panel.

For the remaining clusters, even types are identified (sometimes perfectly, sometimes nearly). Yet, this degree of cleanliness is only achieved because the algorithm is allowed to split one and the same type by the experiences they make. The need for a larger number of clusters is evident with altruists. They have been simulated as non-reactive, cooperative, but noisy. This is why choices look very similar in clusters 1—8. The difference results from the experiences an altruist makes. In cluster 2, they are together with a majority of conditional cooperators, but at least one far-sighted free-rider. As all group members are, at least initially, conditionally cooperative, experiences improve in early periods. However, once a free-rider starts cashing in, the conditional cooperators follow suit, which explains the kink in the second part of the time series. In cluster 3, experiences are more extreme, as far-sighted free-riders have a bigger impact. In cluster 4, experiences never reach the top. This pattern results if hump-shaped players or short-sighted free-riders draw down the contribution level. In clusters 5—8, experiences are flat as the influence of far-sighted free-riders is absent. The composition of the remaining types determines the (nearly or perfectly) constant level of the contributions made by the remaining group members.

At the lower end, clusters 31—34 are also pure. The contributions of short-sighted free-riders are at or near to 0 throughout. But the algorithm needs multiple clusters as experiences differ. Yet, as 3 of the 5 types are themselves reactive, the experience patterns look very different from the experiences that altruists are making. In the most favorable cluster 31, the remaining members are sufficiently cooperative themselves to tolerate exploitation by a single free-rider. By contrast, in clusters 32 and 33, the presence of the free-rider induces cooperative types to reduce their contributions gradually. In cluster 34, a small fraction of conditionally cooperative types is quickly deterred by the prevailing degree of exploitation.

Interestingly, there are also pure clusters of reactive types. In clusters 10 and 11, all players are conditional cooperators. In cluster 10, experiences are initially positive, but deteriorate in the middle of the time series, due to the presence of short-sighted free-riders. In cluster 11, there are no altruists, which is why experiences and contributions never reach the top. But there are no free-riders either, which is why cooperation is stable at an intermediate level.

The algorithm finds even more clusters in which all participants are themselves far-sighted free-riders. There is always the kink in the middle of the series. Clusters 14—17 differ by the experiences these far-sighted free-riders are making. In clusters 15 and 17, these experiences are fairly stable, which must result from the fact that groups are mostly exclusively composed of non-reactive types. In the remaining groups, at least some group members react to the fact that the far-sighted free-rider reduces contributions, by lowering contributions themselves.

In clusters 20—22, only hump-shaped players are to be found. The most characteristic pattern is cluster 20. As long as experiences are very good, hump-shaped players reduce their own contributions. But if far-sighted free-riders start exploiting, experiences fall below the threshold, and hump-shaped players begin stabilizing cooperation. By contrast, in clusters 21 and 22, the contributions of hump-shaped participants stay below the more favorable level of experiences they are making.

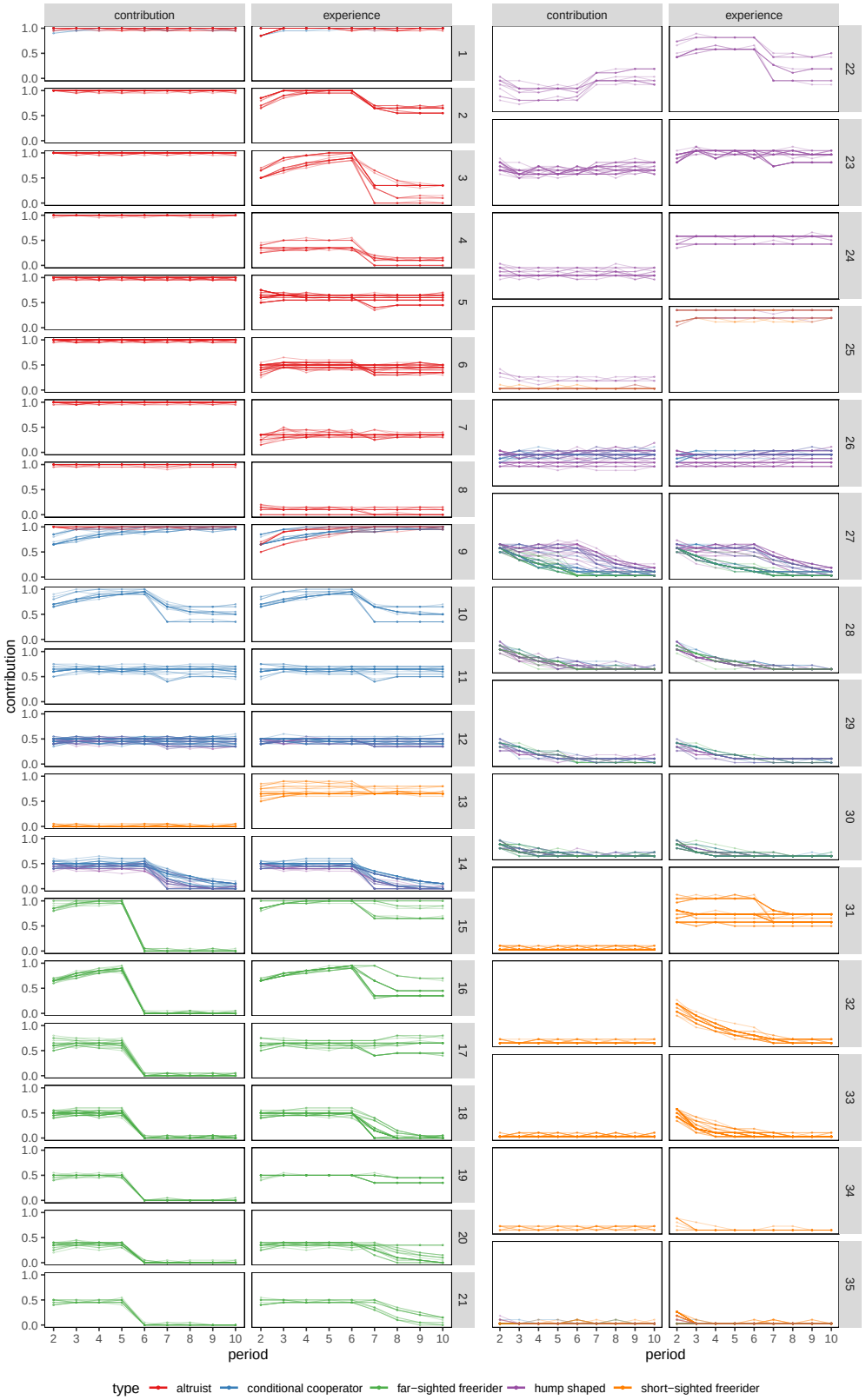


Figure B.1: Exemplary Partitioning of the Simulated Dataset into 35 Clusters for 5 Types

B.2 Details on the Experimental Datasets

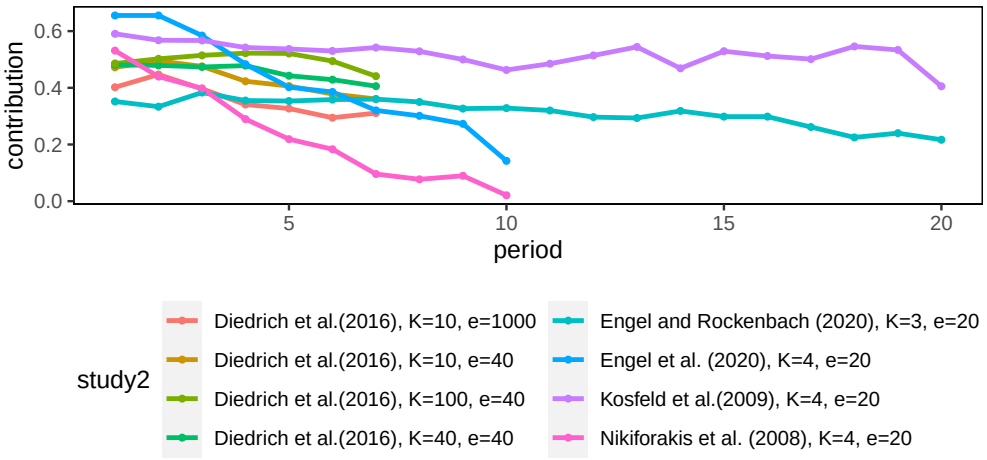


Figure B.2: Means of Participants' Contributions by Study

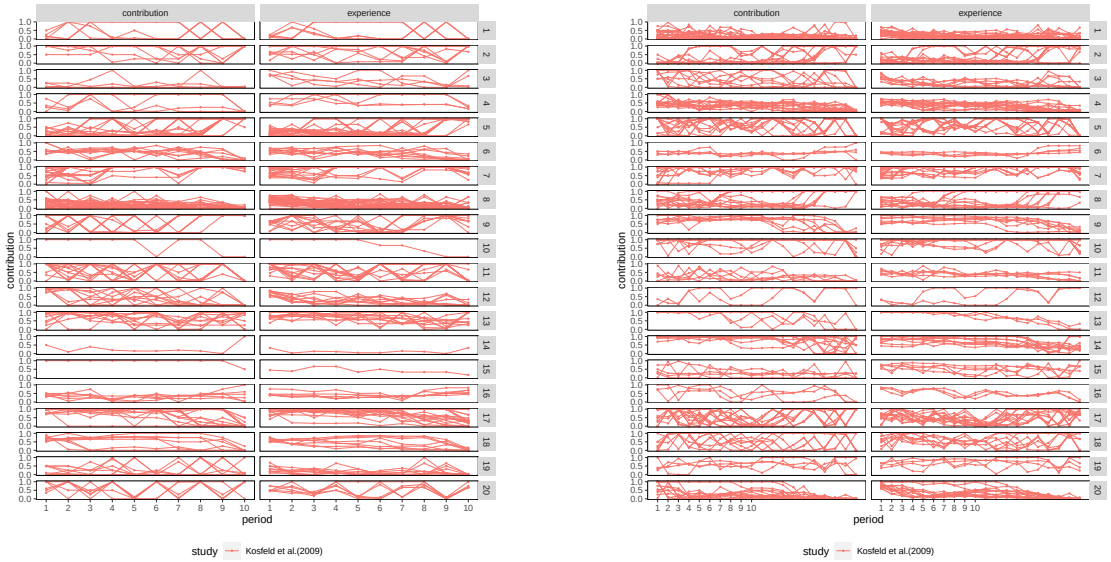
Diederich et al. (2016) play a standard PGG, varying the group size. We include only their treatment 1, which groups 10 individuals at a time. One specialty of their design is long-term rounds, each of exactly 72 hours. Subjects could freely decide when to participate and submit their decision via a device of their choice connected to the internet.

Engel and Rockenbach (2020) add passive bystanders who are either just present, or negative (positively) affected by the contributions active players make to a standard linear public good, to find out why conditional cooperators overreact to negative experiences. We use data from an additional baseline (not reported in the working paper) with no bystanders.

Engel, Kube, et al. (2021) test, in a linear public good, whether selective information about the choices that unrelated third parties have made in the otherwise identical game can increase or decrease contributions. We use data from the baseline, with no manipulation of first impressions.

Kosfeld et al. (2009) play an institution formation game followed by a PGG: Each player decides whether she wants to participate in an organization. Subsequently, players simultaneously determine the number of their contributions to the public good.

Nikiforakis and Normann (2008) deploy a standard PGG and add treatments with punishment options. In our dataset, we only include their control treatment, the PGG without punishment.



(a) Interpolated, 10 periods

(b) Not Interpolated, 20 periods

Figure B.3: Separating Datasets by Periods is Crucial

Table B.1: Subsets by Period

Subset	Periods	Group Size	Subjects
1	10	4, 3	482
2	20	4, 3	362
3	7	10, 40, 100	1210

B.3 Internal Cluster Validation Indices

In Figure B.4, all indices are normalized to the unit interval. Indices to be minimized are recoded and reported as inverse.

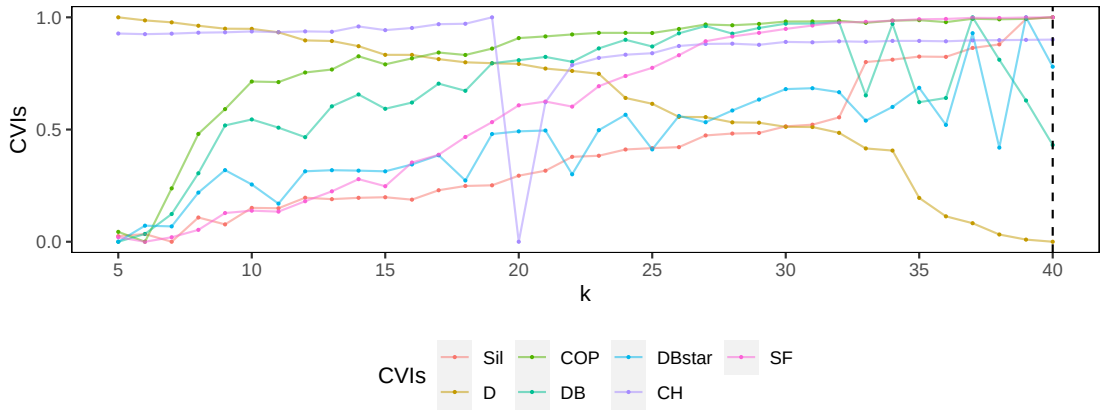


Figure B.4: Simulated Data: Internal Cluster Validation Indices

The dashed vertical line represents the pick based on the rankvote

APPENDIX C

APPENDIX – CHAPTER 3

C.1 Replication

Table C.1: Overview of all Tables and Figures in Landes and Posner (2009) dealing with the Circuit Courts

	Analysis of Court of Appeals Voting: 1925 - 2002
Table 11	Court of Appeals Votes by Subject Matter and Ideology for 538 Court of Appeals Judges Only: 1925 - 2002
Figure 3	Total Votes by Year Appointed to the Court of Appeals
Table 12	Fraction of Mixed (M), Conservative (C) and Liberal (L) Votes for 538 U.S. Court of Appeals Judges by President at Time of Appointment: 1925 - 2002
Table 13	Regression Analysis of Court of Appeals Votes: 1925 - 2002 (t-statistics in parentheses)
Table 14	Regression Analysis of Court of Appeals Votes: 1960 - 2002 (t-statistics in parentheses)
Table 15	Circuit Effects on Ideology of Judges' Votes
Table 16	Regression Analysis of Appellate Court Votes: Current Judges (t-statistics in parentheses)

C.2 Data Pre-processing

We applied pre-processing tailored to our data. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single newline character. Special characters such as 'newline'-characters and roman numbers were removed.

If a potential heading was found within the text, we excluded it. The reason being that such a heading would potentially include biasing information such as judge names. It is especially important to exclude those, as the model could focus on judge names as a proxy for the directionality as most cases were decided without dissent. This is an issue in our empirical context because we would like to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions trying to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part which by its nature goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to an overall worse performance. If we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as 'dissent'. We excluded those entries and did not use them as input.

C.3 All Classifier Input combinations

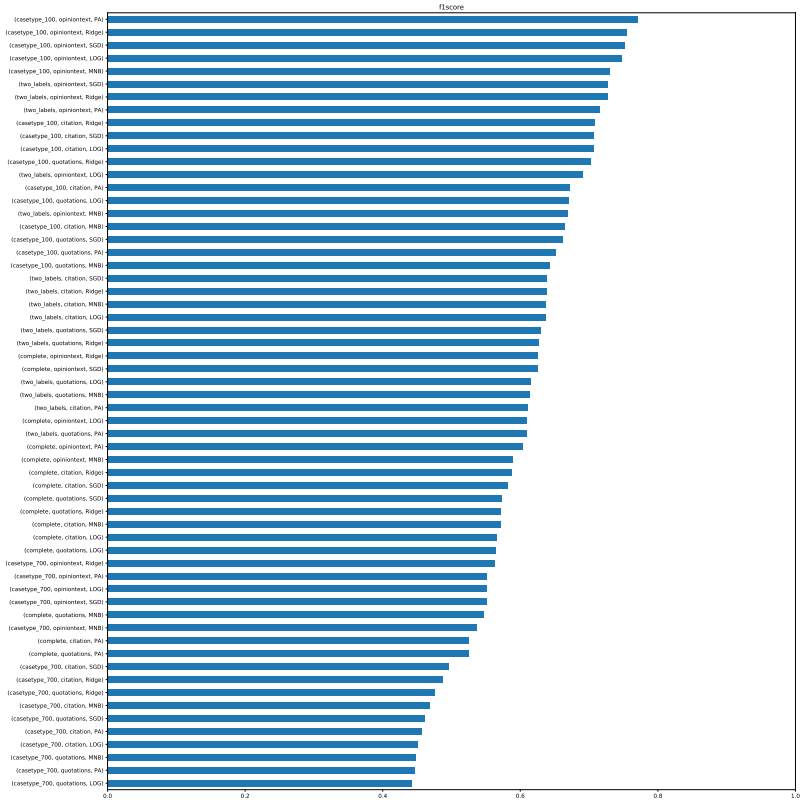
C.4 Judges

Tables C.2, C.3, C.4 and C.5 present yet another way how to assess the performance of the best classifier. We predict the directionality of an opinion and use it to calculate the fraction of conservative or liberal votes by a judge. We split the population of judges by the party of the appointing president, resulting in four different specifications. Overall, actual and predicted fractions of votes by the ten highest ranked judge by the specification are pretty similar and reassures that our classifier performs sufficiently well for our analysis.

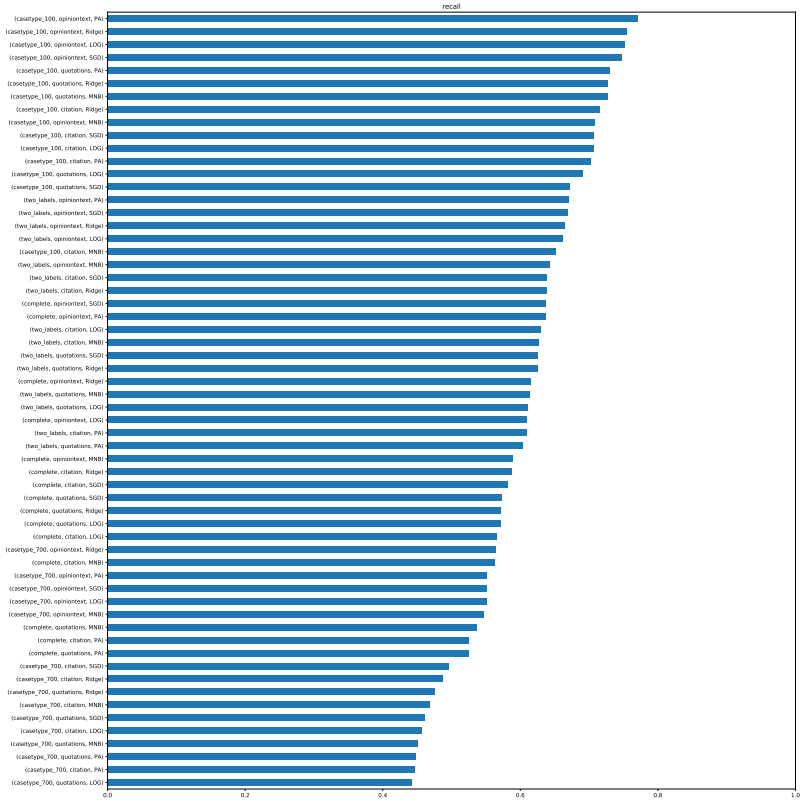
Table C.2: 10 judges with highest fraction of conservative votes, appointed by conservative presidents

Frac con	sum	name
0.89	48	Barksdale, Rhesa H.
0.85	69	Loken, James B.
0.84	65	Hansen, David R.
0.83	110	Easterbrook, Frank H.
0.82	28	O'Scannlain, Diaruid F.
0.82	61	Luttig, J. Michael
0.80	93	Edmondson, James L.
0.80	72	Magill, Frank J.
0.80	104	Boudin, Michael
0.80	45	DeMoss, Harold R., Jr.
Note:		hand-labelled data

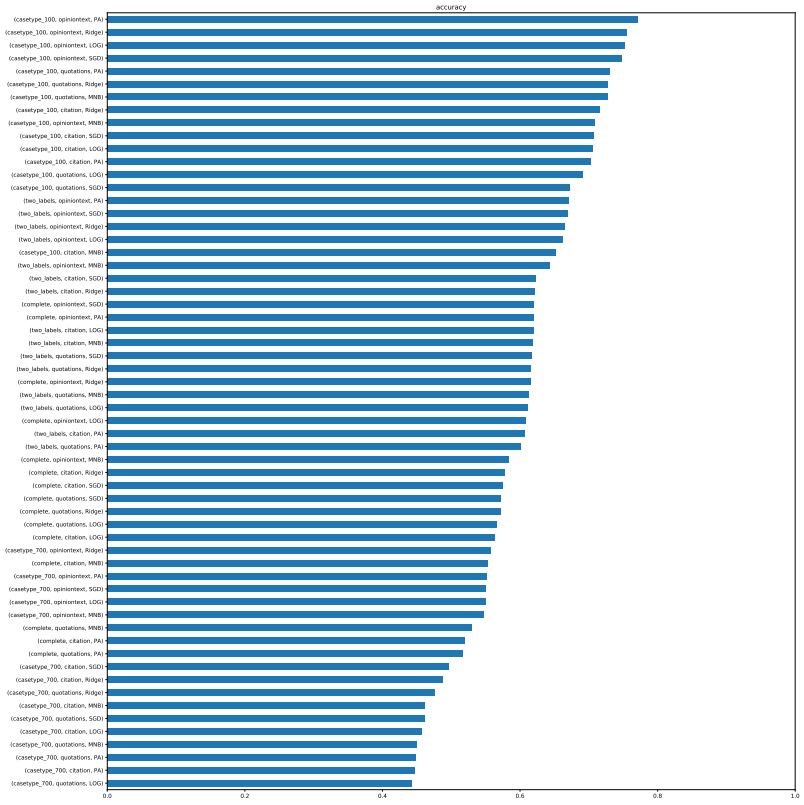
Frac con	sum	name
0.87	48	Barksdale, Rhesa H.
0.87	69	Loken, James B.
0.83	66	Arnold, Morris S.
0.82	109	Easterbrook, Frank H.
0.80	15	Lewis, Robert E.
0.80	65	Hansen, David R.
0.80	44	DeMoss, Harold R., Jr.
0.79	61	Jones, Edith H.
0.79	103	Boudin, Michael
0.78	97	Higginbotham, Patrick E.
Note:		predicted data



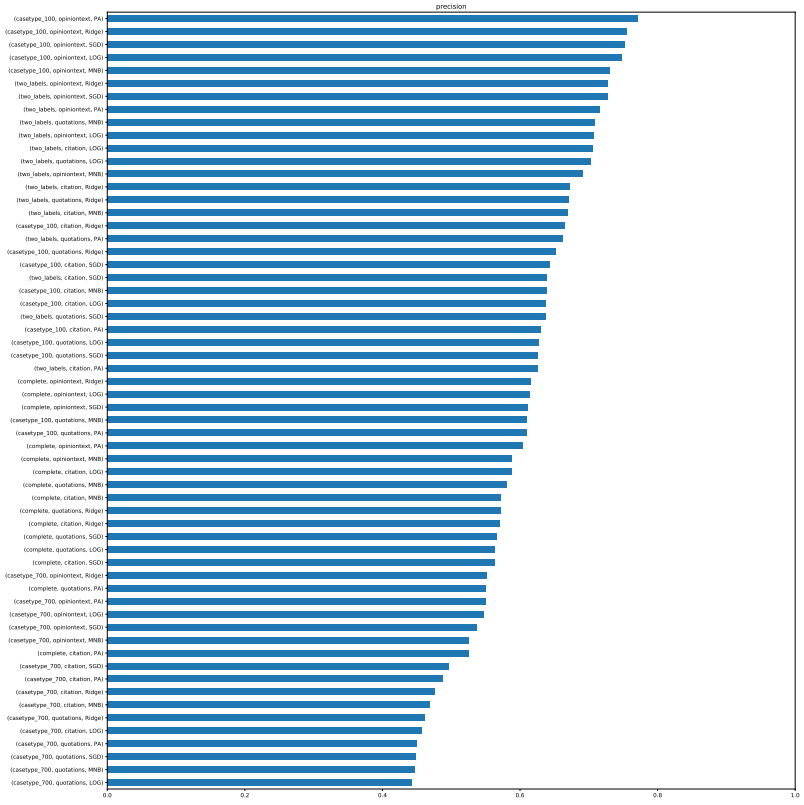
(a) F1 score



(b) Recall score



(c) Accuracy score



(d) Precision score

Figure C.1: Various performance metrics for all different combinations tested

Table C.3: 10 judges with highest fraction of liberal votes, appointed by conservative presidents

Frac lib	sum	name
0.71	11	Thomas, Clarence
0.63	44	Hitz, William
0.59	137	Gibbons, John J.
0.58	39	Waddill, Edmund, Jr.
0.58	46	Miller, William Ernest
0.58	73	Mansmann, Carol Los
0.58	43	Pratt, George C.
0.56	56	Roth, Jane R.
0.56	142	Northcutt, Elliott
0.56	107	Lively, Frederick P.
Note:		hand-labelled data

Frac lib	sum	name
0.63	44	Hitz, William
0.62	11	Thomas, Clarence
0.57	119	Wilbur, Curtis D.
0.56	116	Van Orsdel, Josiah A.
0.56	70	Thompson, Joseph W.
0.56	46	Miller, William Ernest
0.56	55	Roth, Jane R.
0.56	142	Northcutt, Elliott
0.56	108	Lively, Frederick P.
0.55	43	Pratt, George C.
Note:		predicted data

Table C.4: 10 judges with highest fraction of conservative votes, appointed by liberal presidents

Frac con	sum	name
0.89	45	Evans, Terence Thomas
0.84	38	Parker, Robert Manley
0.78	69	Williams, Jerre S.
0.76	83	Garza, Reynaldo
0.75	60	Anderson, Robert P.
0.74	27	King, Carolyn Dineen
0.74	78	Mehaffy, Pat
0.73	131	Miller, Wilbur K., Jr.
0.73	37	Murphy, Michael R.
0.73	11	Kravitch, Phyllis A.
Note:		hand-labelled data

Frac con	sum	name
0.82	44	Evans, Terence Thomas
0.81	37	Parker, Robert Manley
0.80	20	Rutledge, Wiley Blount
0.78	27	King, Carolyn Dineen
0.76	82	Garza, Reynaldo
0.75	134	Breyer, Stephen G.
0.74	163	McMillian, Theodore
0.74	19	Cole, Ransey Guy, Jr.
0.74	68	Williams, Jerre S.
0.73	30	Stewart, Carl Edmond
Note:		predicted data

C.5 Robustness Checks

Additionally to the histograms found in Figure 3.9, we go on to analyze the EBA's statistics on civil cases, displayed by Table C.6a.

For civil cases, we estimated 510 regression models. Figure 3.9a provides information about the share of regression coefficients that are statistically significant as well as lower (column 1) or greater (column 2) than zero. There was no coefficient significant for which the size of at least 50 percent of estimated coefficients lies below zero. By contrast, there were three coefficients found to be significant while having values larger than zero in at least 50 percent of the estimated models. These were the fraction of republican senators at the point of election (92 percent), the fraction of miscellaneous votes (64 percent) as well as circuit 1 (100 percent). Consequently, Leamer (1985)'s EBA (column 3), defines circuit

Table C.5: 10 judges with highest fraction of liberal votes, appointed by liberal presidents

Frac lib	sum	name	Frac lib	sum	name
0.71	11	Faris, Charles	0.66	24	Russell, Robert L.
0.71	11	Thomas, Sidney Runyan	0.63	14	Sarokin, Haddon Lee
0.67	16	Hough, Charles M.	0.63	22	Strum, Louie
0.66	24	Russell, Robert L.	0.62	27	O'Connell, John J.
0.66	51	Haney, Bert E.	0.62	24	Clark, William
0.65	29	Ferguson, Warren J.	0.61	16	Hough, Charles M.
0.63	99	Higginbotham, Aloyisus Leon	0.61	98	Higginbotham, Aloyisus L.
0.63	14	Sarokin, Haddon Lee	0.60	150	Spottswood, Robin. W., III
0.63	22	Strum, Louie	0.60	51	Haney, Bert E.
0.62	24	Clark, William	0.57	31	Lucero, Carlos
Note:		hand-labelled data	*Note:*		predicted data

1 as the only robust variable. Furthermore, Table C.6a includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). Figure 3.9a suggests that a normal distribution does not sufficiently well approximate the regression coefficients' distribution. For this reason, we focus on Sala-i-Martin (1997) EBA results from a model that does make assumptions about the coefficients' distributions. As a rule of thumb, those variables for which more than 90 percent of the regression coefficients' cumulative distribution is located either above or below zero, can be interpreted as being robustly connected with the dependent variable (Hlavac, 2016). For the variables of being black (96 percent), the years of having served as a district court judge (93 percent), as well as for the fraction of economic votes (93 percent), more than 90 percent of the cumulative distributions lie below zero. By contrast, for the variables of being appointed by a conservative president (99 percent), the fraction of miscellaneous votes (98 percent) as well as for circuit 1 (100 percent), more than 90 percent of the cumulative distributions lie above zero.

EBA statistics for criminal cases, displayed in Table C.6b, are interpreted below. Overall, 127 regression models were estimated. Columns 1 and 2 of Table C.6b show the fraction of the respective regression coefficients that are statistically significant and lower or greater than zero at the same time. Only for the dummy variable Black, more than 88 percent of the values estimated were significant and smaller than zero. By contrast, there were three coefficients, Pres (100 percent), circuit 8 (100 percent) and circuit 10 (100 percent) found to be significant and showing more than 50 percent of its values larger than zero. Table C.6b summarizes results from Leamer (1985)'s EBA (column 3). This test concludes that three variables are found to be robustly connected with the dependent variable, which are Pres as well as circuits 8 and 10. Furthermore, Table C.6b includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). As was the case with civil cases, Figure 3.9b suggests that a normal distribution does not fit the coefficients' distribution very well. For this reason, we focus on EBA results from a parameter-free model. For Black (99 percent), more than 90 percent of the cumulative distributions lie

Table C.6: Extreme Bounds Analysis

(a) civil cases

	β sign & < 0	β sign & > 0	leamer robust	cdf $\beta <= 0$ generic	cdf $\beta > 0$ generic
(Intercept)	0.25	0.50	FALSE	0.47	0.53
Pres	0.00	0.92	FALSE	0.01	0.99
SenRep	0.00	0.00	FALSE	0.30	0.70
YrAppt	0.00	0.50	FALSE	0.11	0.89
Gender	0.00	0.00	FALSE	0.33	0.67
Black	0.47	0.00	FALSE	0.96	0.04
DistrictCourt	0.01	0.00	FALSE	0.93	0.07
FracEcon	0.50	0.00	FALSE	0.95	0.05
FracMisc	0.00	0.64	FALSE	0.02	0.98
CircuitVariables1	0.00	1.00	TRUE	0.00	1.00
CircuitVariables2	0.00	0.00	FALSE	0.52	0.48
CircuitVariables3	0.00	0.00	FALSE	0.91	0.09
CircuitVariables4	0.00	0.00	FALSE	0.62	0.38
CircuitVariables5	0.00	0.00	FALSE	0.29	0.71
CircuitVariables6	0.00	0.00	FALSE	0.42	0.58
CircuitVariables7	0.00	0.00	FALSE	0.08	0.92
CircuitVariables8	0.00	0.00	FALSE	0.21	0.79
CircuitVariables9	0.00	0.00	FALSE	0.83	0.17
CircuitVariables10	0.00	0.00	FALSE	0.71	0.29
CircuitVariables11	0.00	0.00	FALSE	0.43	0.57

(b) criminal cases

	β sign & < 0	β sign & > 0	leamer robust	cdf $\beta <= 0$ generic	cdf $\beta > 0$ generic
(Intercept)	0.00	0.50	FALSE	0.14	0.86
Pres	0.00	1.00	TRUE	0.00	1.00
SenRep	0.00	0.00	FALSE	0.70	0.30
YrAppt	0.00	0.00	FALSE	0.44	0.56
Gender	0.00	0.00	FALSE	0.61	0.39
Black	0.88	0.00	FALSE	0.99	0.01
DistrictCourt	0.00	0.00	FALSE	0.78	0.22
CircuitVariables1	0.00	0.00	FALSE	0.06	0.94
CircuitVariables2	0.00	0.00	FALSE	0.60	0.40
CircuitVariables3	0.00	0.00	FALSE	0.71	0.29
CircuitVariables4	0.00	0.00	FALSE	0.46	0.54
CircuitVariables5	0.00	0.00	FALSE	0.25	0.75
CircuitVariables6	0.00	0.00	FALSE	0.49	0.51
CircuitVariables7	0.00	0.00	FALSE	0.07	0.93
CircuitVariables8	0.00	1.00	TRUE	0.01	0.99
CircuitVariables9	0.00	0.00	FALSE	0.33	0.67
CircuitVariables10	0.00	1.00	TRUE	0.01	0.99
CircuitVariables11	0.00	0.00	FALSE	0.14	0.86

below zero. By contrast, for the variables of being appointed by a conservative president (Pres) (100 percent), for circuit 1 (94 percent), circuit 7 (93 percent), circuit 8 (99 percent) and circuit 10 (99 percent) more than 90 percent of the cumulative distributions lie above zero.

APPENDIX D

Appendix – Chapter 4

D.1 Figures

As pointed out in the main text, we modify the COMPAS decile cutoffs slightly as they are calculated against various norm groups - which are unknown to us. The overlap seen in Figure 4.1 gives us evidence that that more than one norm group was used originally. Moreover, it is not even clear whether all individuals' decile scores in the dataset are calculated against the same norm group. Consequently, we the COMPAS deciles used for calculations in the main text are not the original COMPAS ones but the ones normed against the training set. However, in order to show that the renorming impact does not change our overall findings, we include Figure D.1. For the original COMPAS decile scores, the false positives for a decile cutoff greater or equal 4 are 534 (modified: 577) instances while the false negatives would be 93 (modified: 73). For a cutoff 7, the number of false positives is 184 (modified: 202) and 305 false negatives (modified: 275)

In this section, we also include the results for the alternative cutoff-spacing shown in Figure D.3 and Figure D.4. In total, we have 3 possible threshold spacings available. First, the original one by COMPAS imposing a uniform distribution on over the decile scores. Secondly, one that cuts the probability space uniformly, i.e. the upper bound for each decile is +10% risk of recidivism compared to the preceding decile. The first decile would then bin all individuals with an estimated risk of recidivism $\in [0, 10)\%$. Finally, we could do the same for uniform spacing for the estimated COMPAS raw scores. Figure D.2 shows, which values the upper boundaries of the individual deciles would correspond to. To gain predicted risk of recidivism, we again made use of the sigmoid transformation of the raw scores as input. To gain the deciles' raw score boundaries needed to assign the individuals their corresponding decile score, we used the inverse transformation.

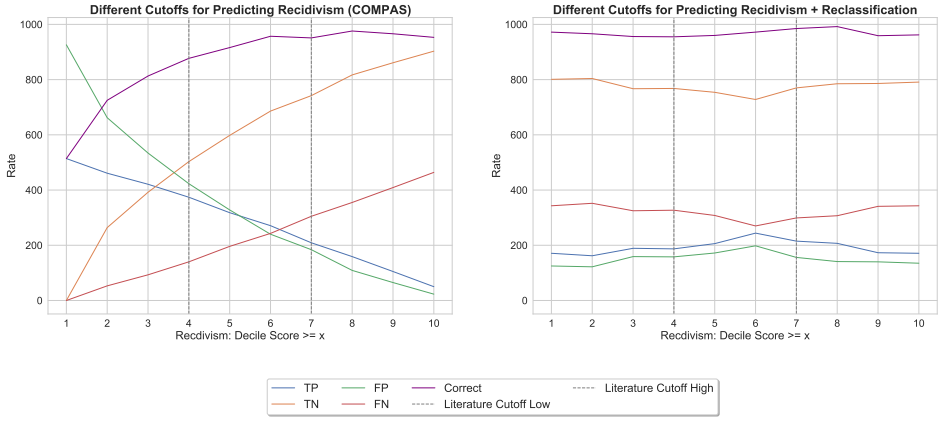


Figure D.1: Comparison COMPAS outcome vs. ex-post correction using the original decile scores.

left panel: original with original COMPAS decile scores, right panel: with ex post correction on original COMPAS decile scores.

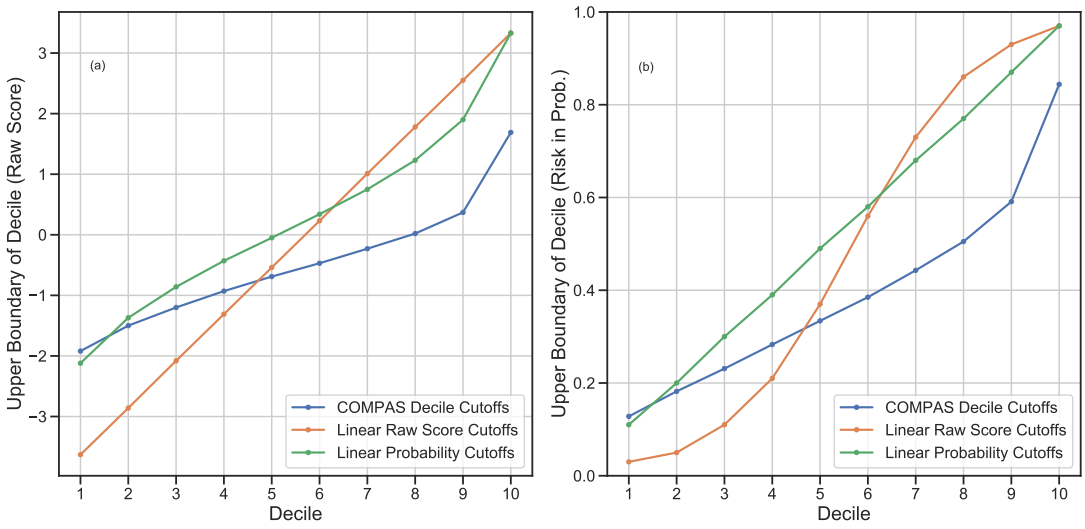


Figure D.2: Different possible spacings for cutoff.

Panel (a) shows the upper raw-score boundaries for each decile, when the respective spacings for the decile generation are applied to the data. Panel (b) shows the same but for the the risk of recidivsim lying between 0-1 (after a sigmoid transformation). Hence they may be interpreted as predicted probability of recidivsim.

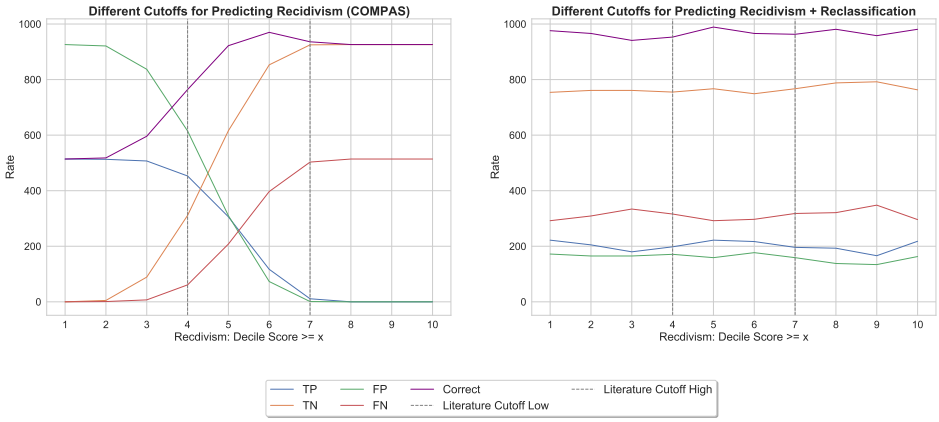


Figure D.3: Outcome of COMPAS and ex-post correction when using when using linear raw score cutoffs.

left panel: original based on COMPAS raw scores, right panel: with ex post correction.

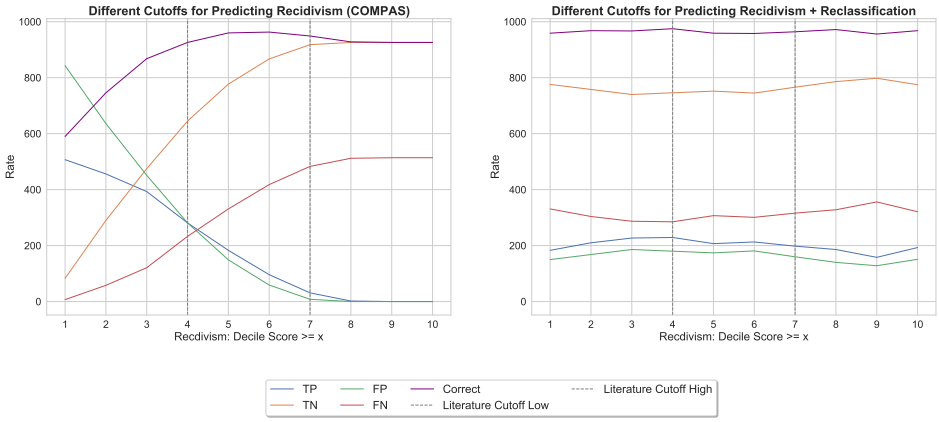


Figure D.4: Outcome of COMPAS and ex-post correction when using when using linear probability cutoffs.

left panel: original based on COMPAS raw scores, right panel: with ex post correction.

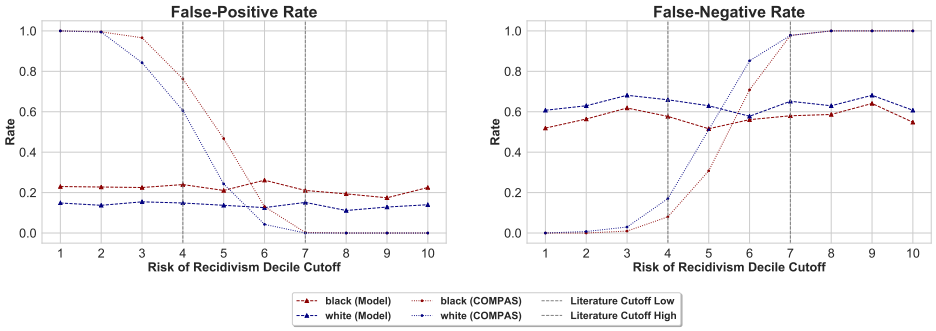


Figure D.5: Racial bias in false positives vs. false negatives when using linear raw score cutoffs.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel).Dotted lines: results when using COMPAS predictions, conditional on cutoff chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Red: black defendants, blue: white defendants. Other races are excluded.

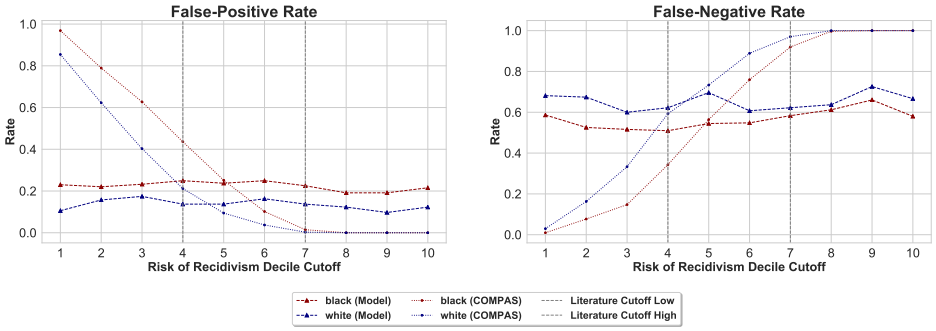


Figure D.6: Racial bias in false positives vs. false negatives when using linear probability cutoffs.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel).Dotted lines: results when using COMPAS predictions, conditional on cutoff chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Red: black defendants, blue: white defendants. Other races are excluded.

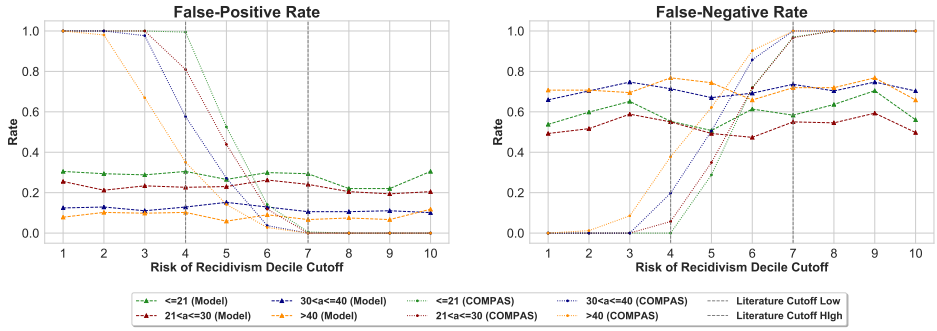


Figure D.7: Age bias in false positives vs. false negatives when using linear raw score cutoffs.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel). Dotted lines: results when using COMPAS predictions, conditional on cutoff chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Green: ≤ 21, red: $(21, 30]$, blue: $(30, 40]$, orange: > 40.

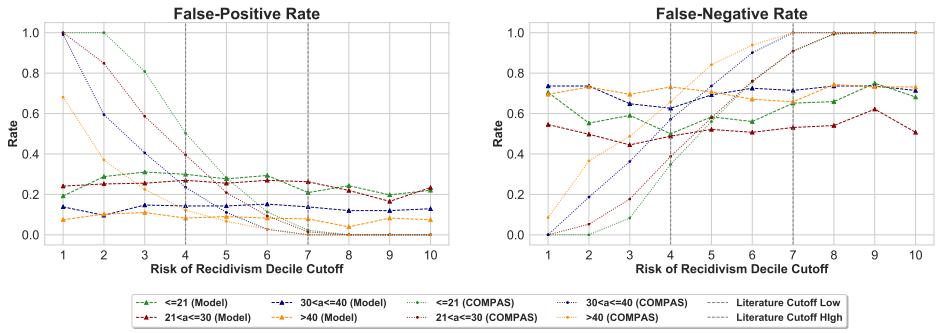


Figure D.8: Age bias in false positives vs. false negatives when using linear probability cutoffs.

The figure shows the rate of defendants incarcerated although they do not recidivate two years after release (left panel) and the rate of defendants released on bail who have recidivated during the next two years (right panel). Dotted lines: results when using COMPAS predictions, conditional on cutoff chosen by the user (x-axis). Dashed lines: results when adding the accuracy correction introduced above. Green: ≤ 21, red: $(21, 30]$, blue: $(30, 40]$, orange: > 40.

D.2 Tables

In order to conduct this analysis, we used two data sources. The first is the raw dataset compiled by ProPublica (Angwin et al., 2013). However, ProPublica's processed dataset is often criticized. While they explain how they calculate individual variables, the supplementary data needed for the calculations is not publicly available. For that reason, Rudin et al. (2020) have gone to tremendous length, to recollect the necessary supplementary data. They made that data available to us upon request. Moreover, they publish the code for generating the final data on their github. Consequently, we used the raw data by Angwin et al. (2013) and the supplementary data as well as the code by Rudin et al. (2020). The latter we adapted slightly to fit our needs, e.g. we constructed the "married"-variable and did not drop all variables from the input data as they did. The exact build of our final dataset, before we applied any preprocessing to it, may be found in Table D.1, Table D.2, Table D.3, Table D.4, and Table D.5

Table D.1: Overview over available input variables – "History of Violence"-subscale items

Feature Name	Type	Values	Explanation
p_juv_fel_count	Integer	count	Prior number of felonies committed by person while the individual was still juvenile
p_felprop_violarrest	Integer	count	Prior violent felony property offense arrests
p_murder_arrest	Integer	count	Prior voluntary manslaughter/murder arrests
p_felassault_arrest	Integer	count	Prior felony assault offense arrests (excluding murder, sex, or domestic violence)
p_misdemassault_arrest	Integer	count	Prior misdemeanor assault offense arrests (excluding murder, sex, domestic violence)
p_famviol_arrest	Integer	count	Prior family violence arrests
p_sex_arrest	Integer	count	Prior misdemeanor assault offense arrests (excluding murder, sex, domestic violence)
p_famviol_arrest	Integer	count	Prior family violence arrests
p_weapons_arrest	Integer	count	Prior weapons offense arrest History of Noncompliance Subscale Items

Table D.2: Overview over available input variables – "History of Criminal Involvement"-subscale items

Feature Name	Type	Values	Explanation
p_charge	Integer	Count	Prior number of charges
p_arrest	Integer	Count	Prior number of arrests
p_jail30	Integer	Count	Prior number of times sentenced to jail 30 days or more
p_prison30	Integer	Count	Prior number of times sentenced to prison 30 days or more
p_prison	Integer	Count	Prior number of times sentenced to prison
p_probation	Integer	Count	Prior number of times sentenced to probation as an adult
is_misdem	Integer	[0,1]	If all charges connected to the current offenses are only misdemeanors = 1, otherwise 0 (i.e. at least one charge is in regards to a felony)

Table D.3: Overview over available input variables – "History of Noncompliance"-subscale items

Feature Name	Type	Values	Explanation
p_n_on_probation	Integer	Count	Prior number of offenses while on probation
p_current_on_probation	Boolean	[0,1]	Current offense committed while on probation
p_prob_revoke	Integer	Count	Number of times probation terms were violeted or probation was revoked

Table D.4: Overview over available input variables – ťťCharacteristics”

Feature Name	Type	Values	Explanation
uid	String	–	Unique identifier; Concatination of id and screening date
first_offense_date	String	–	Date of first offense commited
current_offense_date	String	–	Date of the current offense in question for which COMPAS screening took place
offenses_within_30	Integer	Count	Count all offenses that occured up unitl 30 days prior to screening date
p_felony_count_person	Integer	count	Prior number of felonies committed by person
p_misdem_count_person	Integer	count	Prior number of misdemeanours committed by person
p_charge_violent	Integer	Count	Number of charges against individual falling under violent crimes/offenses
p_current_age	Integer	Age	Age in years of the individual when committing the offense
p_age_first_offense	Integer	Age	Age when committing the first offense (static)
is_married	Boolean	[0,1]	baseline is “single”
is_divorced	Boolean	[0,1]	Baseline is “single”
is_widowed	Boolean	[0,1]	Baseline is “single”
is_separated	Boolean	[0,1]	Baseline is “single”
is_sig_other	Boolean	[0,1]	Baseline is “single”
is_marit_unknown	Boolean	[0,1]	Baseline is “single”
sex	string	[Female, Male]	Gender
race_black	Integer	[0,1]	Individual is black = 1 (baseline is race_other)
race_white	Integer	[0,1]	Individual is white = 1 (baseline is race_other)
race_hispanic	Integer	[0,1]	Individual is hispanic = 1 (baseline is race_other)
race_asian	Integer	[0,1]	Individual is asian = 1 (baseline is race_other)
race_native	Integer	[0,1]	Individual is native = 1 (baseline is race_other)
crim_inv_arrest	Integer	Count	“Criminal Involvement"-scale calculated from features (using arrests) as outlined. Scale is a simple sum of count-based-features. Uses p_charge
crim_inv_charge	Integer	Count	“Criminal Involvement"-scale calculated from features (using charges) as outlined. Scale is a simple sum of Count-based features. Uses p_arrest
vio_hist	Integer	Count	“History of Violence"-scale calculated from features as outlined. Scale is simple sum of count-based features
history_noncomp	Integer	Count	“History of Noncompliance"-scale calculated from features as outlined. Scale is simple sum of count-based features

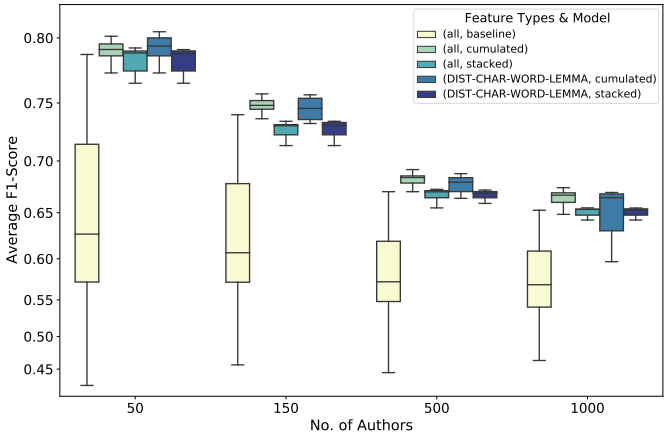
Table D.5: Overview over target characteristics

Feature Name	Type	Values	Explanation
Risk of Failure to Appear_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Failure to Appear_decile_score	Integer	1-10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender as well as race groups.
Risk of Failure to Appear_raw_score	Integer	11-48	COMPAS score Failure to appear
Risk of Recidivism_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Recidivism_decile_score	Integer	1-10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender as well as race groups.
Risk of Recidivism_raw_score	Double	(-3) - (2.36)	COMPAS score “Risk of Recidivism”
Risk of Violence_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Violence_decile_score	Integer	1-10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender
as well as race groups. Risk of Violence_raw_score	Double	(-4.63) - (0.5)	COMPAS score “Risk of Violence”
recid	Integer	[0,1]	Individual is recidivist within two years after screening = 1
recid_violent	Integer	[0,1]	Individual is violent recidivist within two years after screening = 1
recid_proPub	Integer	[0,1]	Individual is recidivist within two years after screening = 1 as calculated by ProPublica
recid_violent_proPub	Integer	[0,1]	Individual is violent recidivist within two years after screening = 1 as calculated by ProPublica

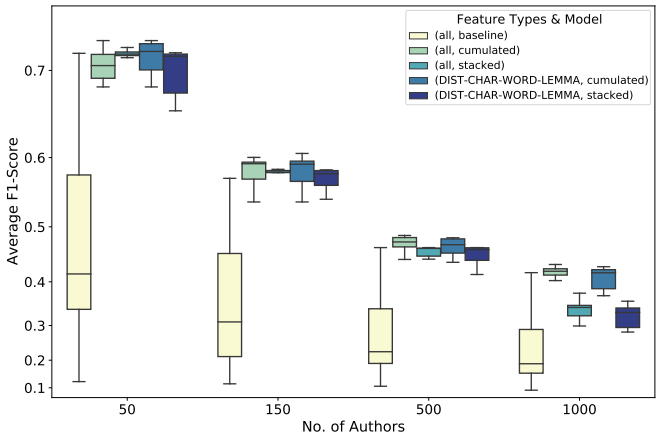
APPENDIX – CHAPTER 5

E.1 Figures

Aggregate Overview



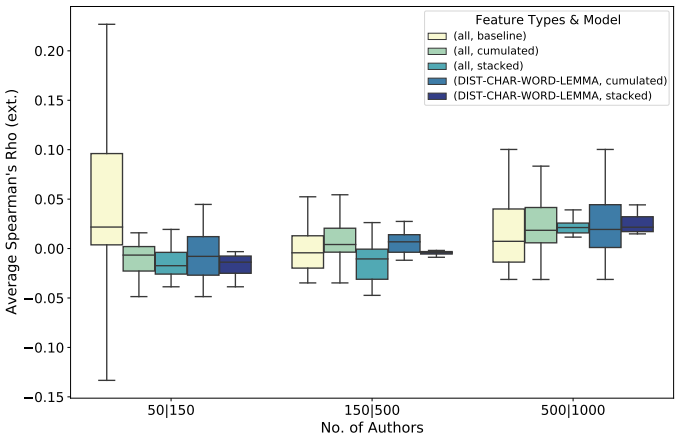
(a) Results for target *gender*



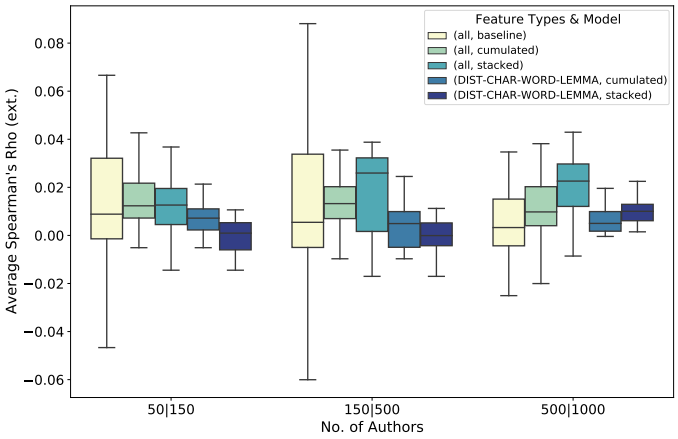
(b) Results for target *age*

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure E.1: F1-Score for all feature type-sets for an input instance length of 100 characters.



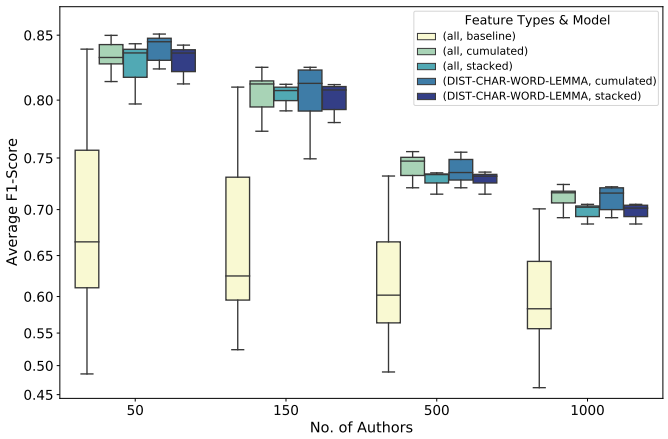
(a) Results for target *gender*



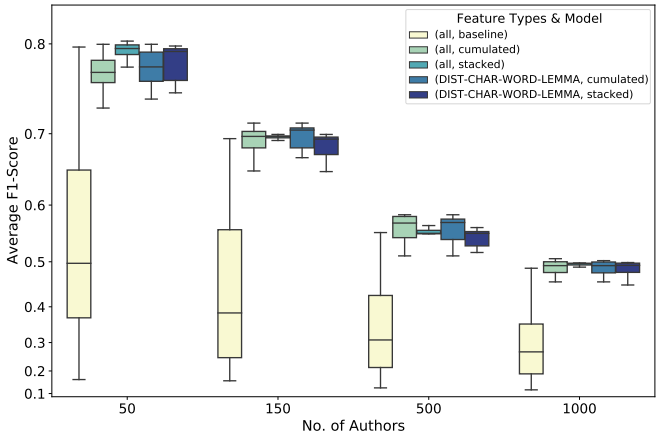
(b) Results for target *age*

Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure E.2: Extended Spearman correlations for all feature type-sets for an input instance length of 100 characters.

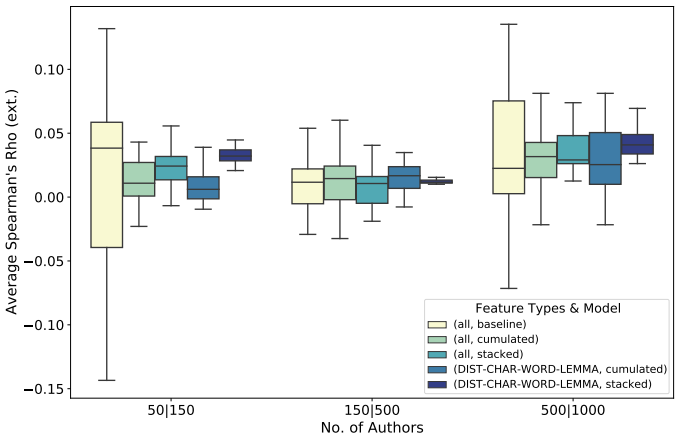


(a) Results for target *gender*

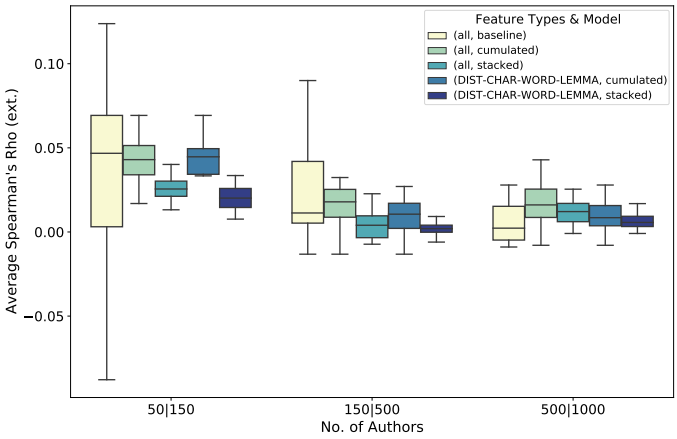


(b) Results for target *age*

Figure E.3: F1-Score for all feature type-sets for an input instance length of 250 characters



(a) Results for target *gender*

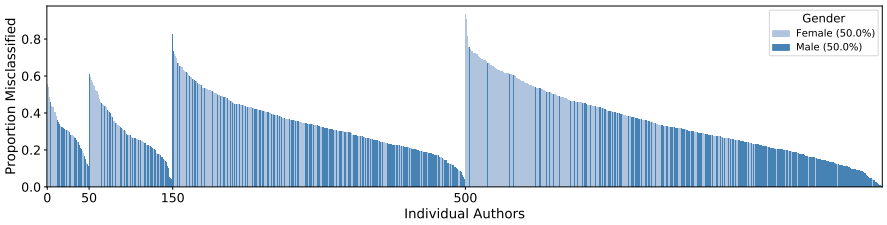


(b) Results for target *age*

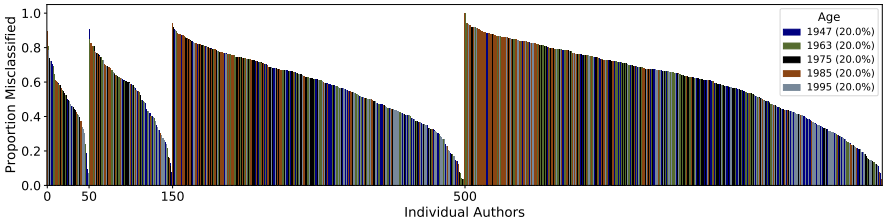
Notes: The figure shows the boxplots for the extended ρ of all models estimated for a given combination of feature types used and way of input, i.e., baseline, cumulated, or stacked.

Figure E.4: Extended Spearman correlation for all feature type sets for an input instance length of 250 characters.

Author-Level Analysis



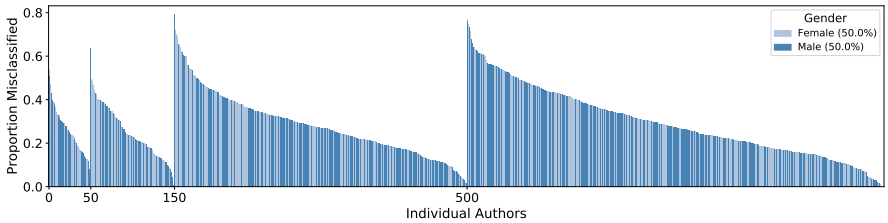
(a) Author-level errors for target *gender*.



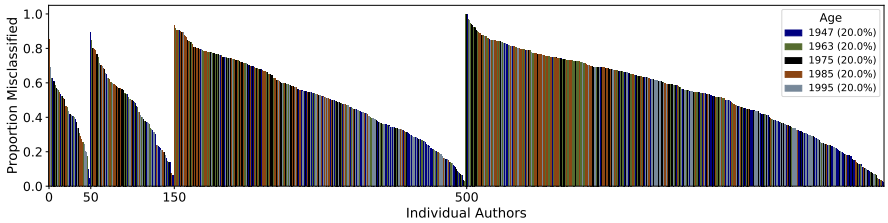
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure E.5: Author-Level Results for the Full feature set with an input instance length of 100 characters.



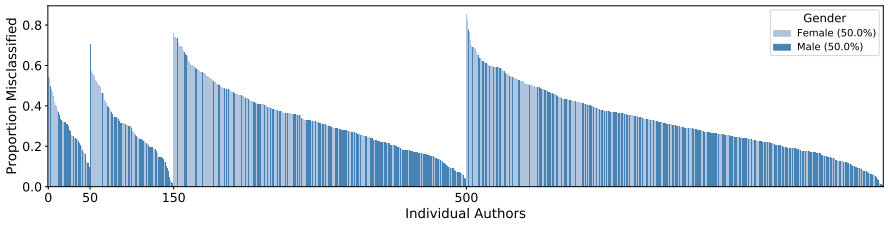
(a) Author-level errors for target *gender*.



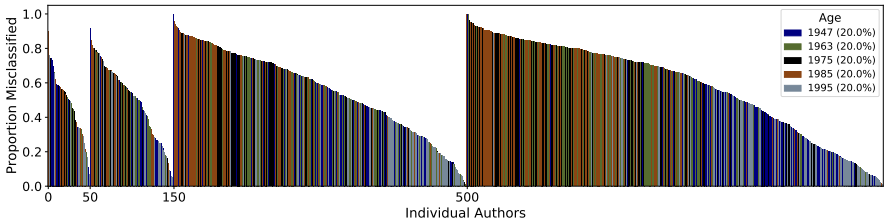
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure E.6: Author-Level Results for the Full feature set in an input instance length of 100 characters.



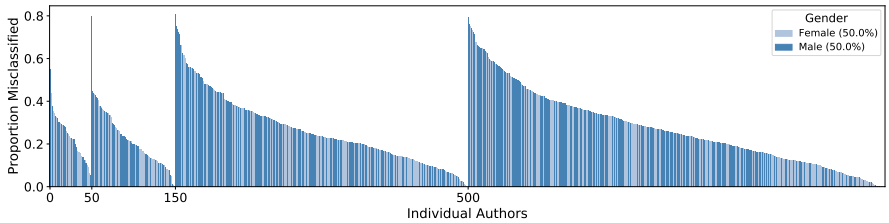
(a) Author-level errors for target *gender*.



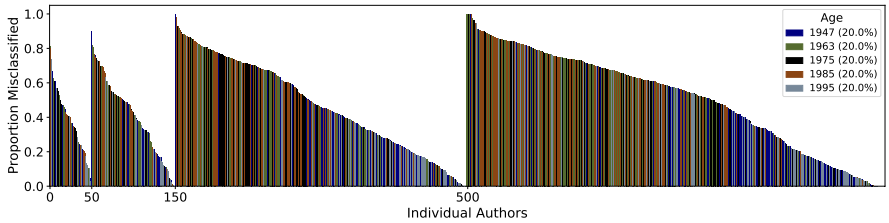
(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure E.7: Author-Level Results for the full feature set with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.



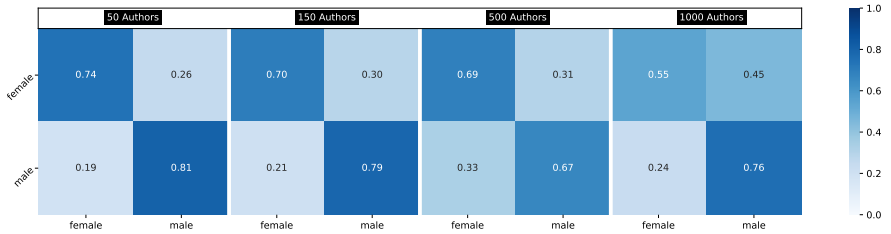
(a) Author-level errors for target *gender*.



(b) Author-level errors for target *age*.

Notes: The figure shows the results when using the full feature set as cumulated input. Each author is a unique instance on the x-axis. The proportion per author is then shown as the y-value. The authors are sorted by their appearance in the respective subsets (i.e., 50, 150, 500, 1000) and according to the proportion of errors within those subsets. The result per author shows the result over all subsets.

Figure E.8: Author-Level Results for the full feature set in an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.

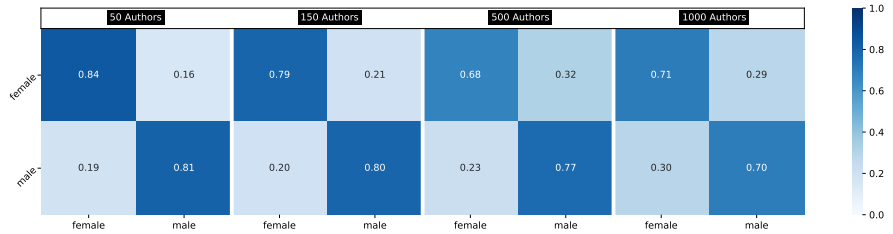


(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.9: Confusion matrices for target *gender* with an input instance length 100 characters - all feature types.

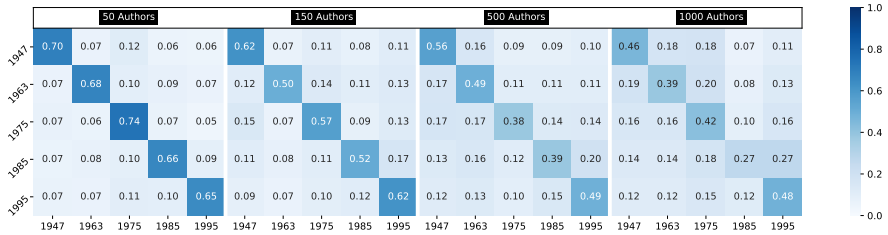


(a) All authors (row-wise normalization).

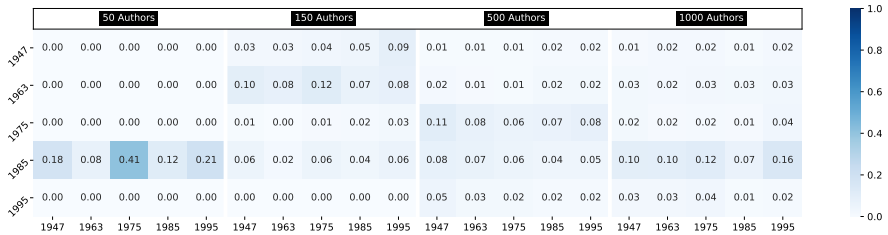
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.10: Confusion matrices for target *gender* with an input instance length of 250 characters - all feature types.



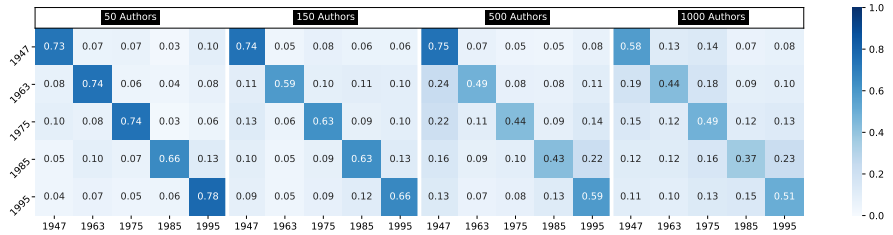
(a) All authors (row-wise normalization).



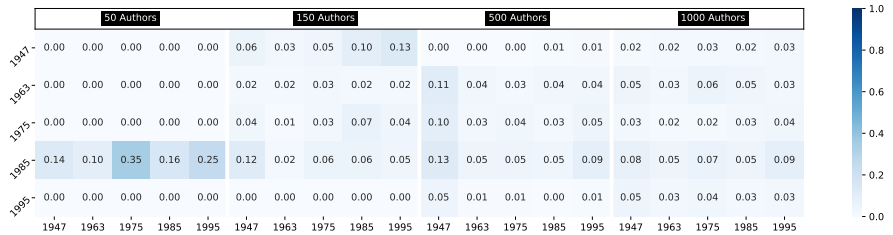
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.11: Confusion matrices for target *age* with an input instance length 100 characters - all feature types.



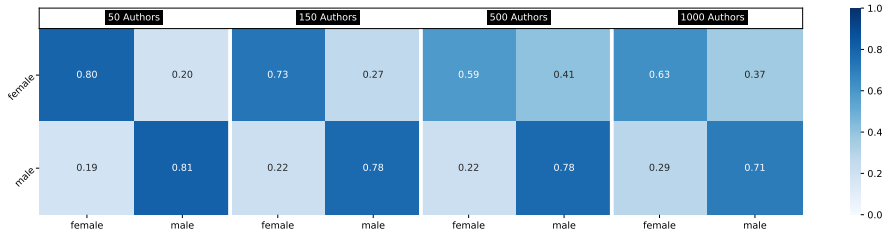
(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the full feature set as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.12: Confusion matrices for target *age* with an input instance length of 250 characters - all feature types.

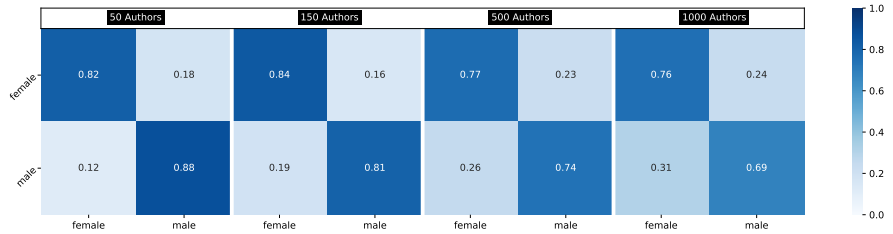


(a) All authors (row-wise normalization).

(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.13: Confusion matrices for target *gender* with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.

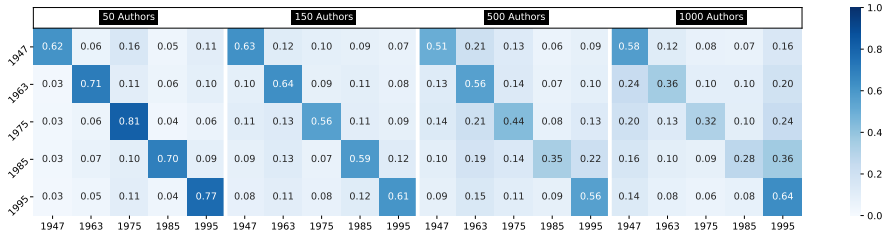


(a) All authors (row-wise normalization).

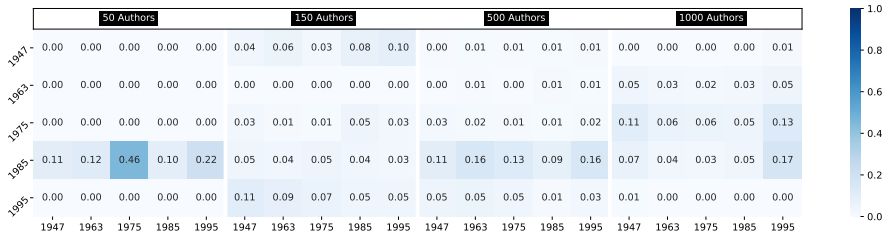
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.14: Confusion matrices for target *gender* with an input instance length of 250 characters - ASIS-CHAR-LEMMA-WORD.



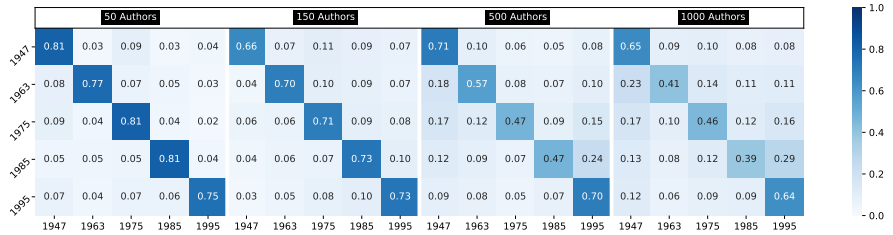
(a) All authors (row-wise normalization).



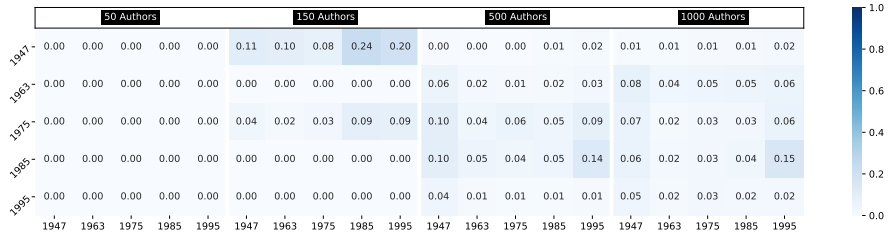
(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 100 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.15: Confusion matrices for target *age* with an input instance length of 100 characters - ASIS-CHAR-LEMMA-WORD.



(a) All authors (row-wise normalization).



(b) Below random guess accuracy (matrix-wise normalization).

Notes: The figure shows confusion matrices for the results produced by using the teh feature types ASIS, CHAR, LEMMA, WORD as cumulated input on an input instance length of 250 characters. The matrix for the respective set of authors is calculated by looking at the respective set in isolation.

Figure E.16: Confusion matrices for target *age* with an input instance length of 250 characters - ASIS-CHAR-LEMMA-WORD.

E.2 Tables

Dataset Statistics

Table E.1: Statistics of the Dataset

		avg_instance		avg_tweet		avg_tweet_per_instance	
	Target	age	gender	age	gender	age	gender
No. of Characters	No. of Authors						
100	50	159.94	162.93	109.21	111.71	1.46	1.46
	150	160.56	162.01	107.39	112.88	1.50	1.44
	500	160.83	161.64	109.30	111.84	1.47	1.45
	1000	160.04	160.92	109.26	111.80	1.46	1.44
250	50	313.46	315.85	109.39	112.14	2.87	2.82
	150	313.21	315.31	107.38	112.75	2.92	2.80
	500	313.21	314.61	109.06	111.62	2.87	2.82
	1000	313.33	314.29	109.23	111.81	2.87	2.81
500	50	565.52	568.76	108.88	112.09	5.19	5.07
	150	565.67	568.60	107.37	112.89	5.27	5.04
	500	566.35	567.42	109.17	111.80	5.19	5.08
	1000	566.01	567.42	109.16	111.92	5.19	5.07

Table E.2: Statistics of the Dataset

		avg_instance		avg_tweet		avg_tweet_per_instance	
	Target	age	gender	age	gender	age	gender
No. of Characters	No. of Authors						
100	50	161.16	163.39	110.30	112.38	1.46	1.45
	150	160.11	162.66	107.11	112.99	1.49	1.44
	500	160.73	161.51	109.04	111.86	1.47	1.44
	1000	160.10	161.06	109.31	111.83	1.46	1.44
250	50	312.62	316.00	108.98	112.14	2.87	2.82
	150	313.13	316.26	107.63	113.00	2.91	2.80
	500	313.25	314.76	109.22	111.64	2.87	2.82
	1000	313.29	314.28	109.07	111.87	2.87	2.81
500	50	565.07	568.36	109.31	110.78	5.17	5.13
	150	566.11	568.92	107.51	112.82	5.27	5.04
	500	565.73	567.60	109.05	111.69	5.19	5.08
	1000	566.23	567.34	109.21	111.85	5.18	5.07

Table E.3: Statistics of the Dataset

		avg_instance		avg_tweet		avg_tweet_per_instance	
	Target	age	gender	age	gender	age	gender
No. of Characters	No. of Authors						
100	50	159.93	162.32	109.42	111.75	1.46	1.45
	150	161.05	162.54	107.75	112.95	1.49	1.44
	500	160.74	161.81	109.27	111.92	1.47	1.45
	1000	160.09	161.03	109.28	111.96	1.47	1.44
250	50	313.19	315.01	108.30	111.55	2.89	2.82
	150	313.55	315.26	107.24	112.85	2.92	2.79
	500	313.39	314.64	108.98	111.86	2.88	2.81
	1000	313.23	314.55	109.23	111.85	2.87	2.81
500	50	566.58	569.37	109.40	112.97	5.18	5.04
	150	566.12	568.69	107.69	112.78	5.26	5.04
	500	566.17	567.78	109.24	111.45	5.18	5.09
	1000	566.14	567.36	109.12	111.69	5.19	5.08

Baseline

Minimum of Characters: 100

Table E.4: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.5886	0.5711	0.5455	0.4567	0.5639	0.5594	0.5387	0.4881
	2-3	0.6071	0.5727	0.5992	0.5881	0.5358	0.4757	0.5549	0.5379
	2-3-4	0.6354	0.6251	0.5693	0.5099	0.5521	0.5267	0.5496	0.5167
	2-3-4-5	0.6451	0.6383	0.6076	0.5975	0.5714	0.5708	0.5554	0.5312
CHAR	2	0.7080	0.7080	0.6725	0.6724	0.6152	0.6152	0.6053	0.6047
	2-3	0.7555	0.7555	0.7127	0.7127	0.6560	0.6560	0.6389	0.6380
	2-3-4	0.7656	0.7656	0.7210	0.7210	0.6624	0.6624	0.6437	0.6423
	2-3-4-5	0.7650	0.7649	0.7201	0.7199	0.6617	0.6617	0.6434	0.6420
ASIS	2	0.7362	0.7360	0.6989	0.6982	0.6331	0.6331	0.6212	0.6212
	2-3	0.7683	0.7643	0.7351	0.7350	0.6662	0.6661	0.6475	0.6462
	2-3-4	0.7799	0.7787	0.7403	0.7403	0.6711	0.6710	0.6526	0.6525
	2-3-4-5	0.7883	0.7878	0.7403	0.7402	0.6706	0.6704	0.6523	0.6523
POS	1	0.5671	0.5450	0.5513	0.5313	0.5446	0.5425	0.5463	0.5451
	1-2	0.5950	0.5867	0.5912	0.5912	0.5622	0.5621	0.5611	0.5584
	1-2-3	0.6086	0.6028	0.6005	0.6003	0.5711	0.5711	0.5697	0.5682
TAG	1	0.5895	0.5714	0.5745	0.5740	0.5500	0.5496	0.5514	0.5498
	1-2	0.6293	0.6291	0.6117	0.6112	0.5744	0.5743	0.5709	0.5700
	1-2-3	0.6372	0.6343	0.6213	0.6213	0.5820	0.5820	0.5787	0.5779
DEP	1	0.5752	0.5612	0.5676	0.5672	0.5404	0.5389	0.5423	0.5413
	1-2	0.5951	0.5926	0.5864	0.5738	0.5569	0.5528	0.5585	0.5553
	1-2-3	0.6070	0.5987	0.6036	0.6025	0.5670	0.5670	0.5645	0.5615
LEMMA	1	0.6719	0.6653	0.6351	0.6273	0.5994	0.5968	0.5901	0.5843
	1-2	0.6958	0.6957	0.6465	0.6463	0.6019	0.5985	0.5930	0.5860
WORD	1	0.6371	0.6258	0.6271	0.6270	0.5865	0.5834	0.5817	0.5760
	1-2	0.6490	0.6485	0.6295	0.6295	0.5905	0.5901	0.5847	0.5809
NUM	1	0.5573	0.5551	0.5437	0.4778	0.5327	0.5300	0.5273	0.4637

Table E.5: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
CHAR	2	0.0011	0.0221	0.0639	0.0061	-0.0048	-0.0542
	2-3	0.0047	-0.0198	-0.0146	-0.0051	0.0206	-0.0095
	2-3-4	-0.0036	-0.0135	0.0033	-0.0017	0.0115	0.0061
	2-3-4-5	0.0150	0.0178	0.0008	-0.0004	0.0053	0.0064
ASIS	2	0.0423	0.0470	0.0020	0.0311	-0.0260	0.0004
	2-3	0.0282	-0.0056	0.0137	0.0174	0.0192	0.0201
	2-3-4	0.0117	0.0059	-0.0096	-0.0011	-0.0056	0.0001
	2-3-4-5	0.0034	0.0093	-0.0170	0.0070	-0.0004	-0.0125
POS	1	0.2719	0.2719	-0.0719	-0.0719	-0.2421	-0.2421
	1-2	0.2268	0.2242	0.0524	0.0709	0.1519	0.1775
	1-2-3	0.1066	0.0676	-0.1519	-0.1720	0.0127	0.0380
TAG	1	-0.0293	-0.0293	0.0516	0.0516	-0.0968	-0.0968
	1-2	0.1570	0.1739	-0.1257	-0.1234	0.0425	0.0401
	1-2-3	-0.0161	-0.0200	0.0004	0.0193	0.0080	0.0402
DEP	1	0.2392	0.2392	-0.0207	-0.0207	0.0386	0.0386
	1-2	0.1057	0.1265	0.0345	0.0221	-0.0154	0.0064
	1-2-3	0.0077	0.0151	-0.0730	-0.0300	0.0747	0.0976
LEMMA	1	0.0673	-0.0261	0.0038	0.0636	-0.0132	0.0167
	1-2	0.0293	0.0076	-0.0149	0.0084	-0.0139	-0.0066
WORD	1	0.0276	0.0771	0.0105	0.0330	0.0067	0.0167
	1-2	0.0128	-0.0301	-0.0266	0.0056	0.0721	0.0288
NUM	1	-0.1333	-0.1333	-0.0167	-0.0167	-0.4833	-0.4833

Table E.6: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Age							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.3648	0.3249	0.2771	0.2136	0.2645	0.2207	0.2114	0.1008
	2-3	0.3689	0.3502	0.3350	0.3266	0.2162	0.1061	0.2082	0.0899
	2-3-4	0.4086	0.3943	0.3516	0.3444	0.2712	0.2120	0.2197	0.1133
	2-3-4-5	0.4309	0.4286	0.3414	0.3188	0.3018	0.2560	0.2789	0.2226
CHAR	2	0.5608	0.5608	0.4483	0.4443	0.3500	0.3291	0.2982	0.2860
	2-3	0.6430	0.6446	0.5326	0.5312	0.4334	0.4298	0.3819	0.3715
	2-3-4	0.6637	0.6642	0.5473	0.5421	0.4485	0.4477	0.4163	0.4118
	2-3-4-5	0.6649	0.6651	0.5529	0.5486	0.4474	0.4439	0.4124	0.4020
ASIS	2	0.6257	0.6254	0.4874	0.4846	0.3811	0.3769	0.3169	0.3009
	2-3	0.7053	0.7051	0.5598	0.5574	0.4483	0.4453	0.3982	0.3974
	2-3-4	0.7173	0.7172	0.5747	0.5714	0.4639	0.4618	0.4147	0.3992
	2-3-4-5	0.6978	0.6987	0.5748	0.5724	0.4673	0.4648	0.4229	0.4183
POS	1	0.2464	0.1722	0.2145	0.1158	0.2197	0.1573	0.2041	0.1218
	1-2	0.3575	0.3446	0.2257	0.1213	0.2553	0.2194	0.2176	0.1414
	1-2-3	0.4003	0.3939	0.2650	0.2222	0.2674	0.2358	0.2331	0.1823
TAG	1	0.2639	0.1977	0.2155	0.1222	0.2349	0.1735	0.2121	0.1724
	1-2	0.4195	0.4177	0.2970	0.2637	0.2743	0.2396	0.2299	0.1810
	1-2-3	0.4483	0.4456	0.3425	0.3341	0.2540	0.2068	0.2437	0.1885
DEP	1	0.3002	0.2681	0.2408	0.2056	0.2166	0.1777	0.2151	0.1850
	1-2	0.3638	0.3521	0.2843	0.2659	0.2329	0.1688	0.2283	0.1882
	1-2-3	0.3957	0.3920	0.3082	0.2939	0.2669	0.2493	0.2419	0.2160
LEMMA	1	0.4560	0.4543	0.3269	0.3092	0.2559	0.2191	0.2504	0.2193
	1-2	0.4872	0.4864	0.3484	0.3368	0.2939	0.2835	0.2367	0.1861
WORD	1	0.4193	0.4139	0.3010	0.2767	0.2441	0.2052	0.2244	0.1605
	1-2	0.4274	0.4249	0.3169	0.3090	0.2569	0.2326	0.2386	0.1985
NUM	1	0.2694	0.2420	0.2263	0.1991	0.2260	0.1932	0.2111	0.2053

Table E.7: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
CHAR	2	0.0052	0.0143	0.0144	-0.0001	0.0159	0.0310
	2-3	0.0018	-0.0006	-0.0138	0.0003	-0.0044	0.0119
	2-3-4	0.0129	-0.0036	0.0038	-0.0071	0.0016	0.0145
	2-3-4-5	0.0123	0.0102	0.0071	0.0085	0.0051	0.0107
ASIS	2	0.0063	0.0106	-0.0010	-0.0015	0.0112	-0.0048
	2-3	0.0066	0.0100	–	-0.0011	-0.0002	0.0086
	2-3-4	-0.0014	0.0033	0.0075	0.0063	0.0072	0.0010
	2-3-4-5	0.0099	0.0034	0.0036	0.0036	0.0101	0.0097
POS	1	-0.0312	-0.0312	0.0375	0.0375	0.1063	0.1063
	1-2	0.0078	0.0058	0.0881	0.0757	0.1056	0.0972
	1-2-3	0.0443	0.0367	0.0603	0.0675	0.1082	0.1032
TAG	1	0.1134	0.1134	-0.0600	-0.0600	-0.0528	-0.0528
	1-2	0.0647	0.0702	0.0694	0.0775	0.0347	0.0494
	1-2-3	0.0339	0.0345	0.0370	0.0329	0.0003	0.0022
DEP	1	0.0666	0.0860	-0.0819	-0.0819	-0.0113	-0.0113
	1-2	0.0529	-0.0187	0.0834	0.0891	0.0540	0.0571
	1-2-3	-0.0014	-0.0147	0.0361	0.0407	0.0603	0.0692
LEMMA	1	0.0114	0.0114	-0.0066	0.0111	-0.0040	-0.0182
	1-2	-0.0089	0.0103	0.0012	0.0139	-0.0251	-0.0304
WORD	1	0.0266	0.0294	0.0270	0.0137	0.0046	-0.0065
	1-2	-0.0109	0.0096	0.0032	-0.0119	-0.0010	0.0006
NUM	1	-0.0467	-0.0467	-0.2500	-0.2500	-0.1267	-0.1267

Minimum of Characters: 250

Table E.8: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.6002	0.5633	0.6026	0.6026	0.5649	0.5558	0.5543	0.5311
	2-3	0.6261	0.6071	0.5801	0.5326	0.5771	0.5764	0.5541	0.5267
	2-3-4	0.6586	0.6551	0.6147	0.5988	0.5624	0.5471	0.5583	0.5314
	2-3-4-5	0.6711	0.6692	0.6203	0.6062	0.5733	0.5666	0.5675	0.5537
CHAR	2	0.7495	0.7480	0.7247	0.7242	0.6605	0.6603	0.6300	0.6218
	2-3	0.8031	0.8020	0.7781	0.7781	0.7125	0.7122	0.6793	0.6774
	2-3-4	0.8087	0.8077	0.7905	0.7903	0.7161	0.7137	0.6895	0.6880
	2-3-4-5	0.8178	0.8178	0.7804	0.7797	0.7210	0.7201	0.6897	0.6886
ASIS	2	0.7844	0.7841	0.7561	0.7556	0.6812	0.6812	0.6587	0.6575
	2-3	0.8256	0.8243	0.7964	0.7958	0.7249	0.7240	0.6905	0.6893
	2-3-4	0.8397	0.8397	0.8076	0.8068	0.7327	0.7320	0.7004	0.6996
	2-3-4-5	0.8383	0.8379	0.8106	0.8105	0.7335	0.7332	0.7013	0.7007
POS	1	0.5967	0.5923	0.5683	0.5643	0.5639	0.5639	0.5613	0.5609
	1-2	0.6294	0.6226	0.6046	0.6020	0.5863	0.5840	0.5729	0.5623
	1-2-3	0.6499	0.6486	0.6162	0.6081	0.6003	0.5997	0.5876	0.5823
TAG	1	0.6276	0.6267	0.5856	0.5853	0.5700	0.5648	0.5643	0.5643
	1-2	0.6619	0.6619	0.6324	0.6306	0.6049	0.6039	0.5879	0.5829
	1-2-3	0.6770	0.6769	0.6562	0.6546	0.6141	0.6139	0.6029	0.6020
DEP	1	0.5978	0.5962	0.5812	0.5811	0.5586	0.5557	0.5572	0.5572
	1-2	0.6192	0.6125	0.6050	0.5998	0.5802	0.5800	0.5709	0.5642
	1-2-3	0.6309	0.6297	0.6249	0.6211	0.5904	0.5875	0.5857	0.5849
LEMMA	1	0.7198	0.7179	0.7046	0.7038	0.6504	0.6479	0.6349	0.6324
	1-2	0.7388	0.7385	0.7137	0.7136	0.6554	0.6554	0.6408	0.6385
WORD	1	0.6966	0.6961	0.6869	0.6869	0.6354	0.6352	0.6294	0.6292
	1-2	0.7011	0.7000	0.6914	0.6912	0.6409	0.6396	0.6313	0.6298
NUM	1	0.5760	0.5527	0.5416	0.5414	0.5478	0.5145	0.5431	0.5431

Table E.9: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
CHAR	2	0.1271	-0.0684	0.0223	0.0369	0.0445	0.0490
	2-3	0.0599	-0.0195	0.0212	0.0012	0.0459	0.0043
	2-3-4	0.0587	-0.0164	0.0135	0.0124	0.0036	0.0127
	2-3-4-5	0.0421	0.0059	0.0037	0.0049	0.0188	0.0229
ASIS	2	0.1042	0.0513	0.0164	0.0274	0.0377	0.0374
	2-3	0.0581	0.0135	0.0202	0.0295	0.0128	-0.0127
	2-3-4	0.0456	0.0216	0.0098	0.0092	0.0194	0.0276
	2-3-4-5	0.0399	0.0271	0.0061	0.0113	0.0221	0.0118
POS	1	-0.0754	-0.0754	0.3930	0.3930	0.4070	0.4070
	1-2	-0.2452	-0.1970	-0.0245	-0.0245	0.2085	0.2085
	1-2-3	-0.0732	-0.1022	-0.0618	-0.0595	-0.0122	-0.0147
TAG	1	-0.0482	0.0273	0.2040	0.2040	0.0438	0.0438
	1-2	0.0165	0.0799	-0.0851	-0.1044	-0.0175	-0.0084
	1-2-3	0.0278	—	-0.0074	0.0033	0.0023	-0.0284
DEP	1	-0.1435	-0.2558	-0.0789	-0.0943	0.2496	0.2496
	1-2	0.0537	-0.0038	-0.0292	-0.0330	-0.0715	-0.0513
	1-2-3	0.0377	-0.0120	-0.1215	-0.0453	0.0575	0.0265
LEMMA	1	-0.0620	-0.0509	0.0051	-0.0123	-0.0430	0.0474
	1-2	-0.0081	0.0508	0.0033	0.0062	-0.0042	0.0166
WORD	1	-0.0846	-0.0162	0.0407	0.0240	0.0229	0.0253
	1-2	-0.0132	0.0485	0.0153	0.0136	0.0064	-0.0079
NUM	1	0.6167	0.6167	0.1333	0.1333	-0.2333	-0.2333

Table E.10: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.3910	0.3670	0.2839	0.2198	0.2414	0.1601	0.2796	0.2179
	2-3	0.4489	0.4159	0.3855	0.3728	0.2272	0.1273	0.3201	0.3105
	2-3-4	0.4944	0.4925	0.4022	0.3925	0.3426	0.3169	0.3057	0.2629
	2-3-4-5	0.4819	0.4712	0.4192	0.4129	0.3373	0.3023	0.3209	0.2863
CHAR	2	0.6413	0.6425	0.5527	0.5512	0.4262	0.4188	0.3499	0.3432
	2-3	0.7199	0.7226	0.6369	0.6363	0.5148	0.5104	0.4514	0.4513
	2-3-4	0.7341	0.7347	0.6701	0.6691	0.5333	0.5311	0.4779	0.4728
	2-3-4-5	0.7422	0.7438	0.6661	0.6656	0.5377	0.5391	0.4825	0.4777
ASIS	2	0.6750	0.6803	0.5870	0.5838	0.4627	0.4534	0.4018	0.3890
	2-3	0.7761	0.7761	0.6804	0.6799	0.5450	0.5430	0.4757	0.4728
	2-3-4	0.7803	0.7813	0.6948	0.6938	0.5562	0.5544	0.4863	0.4742
	2-3-4-5	0.7962	0.7968	0.6916	0.6913	0.5535	0.5516	0.4938	0.4870
POS	1	0.3297	0.2573	0.2708	0.1954	0.2242	0.1283	0.2097	0.1190
	1-2	0.3974	0.3730	0.2731	0.1993	0.2838	0.2495	0.2642	0.2162
	1-2-3	0.4580	0.4464	0.3689	0.3583	0.3194	0.3139	0.2485	0.1735
TAG	1	0.3453	0.3081	0.2867	0.2571	0.2193	0.1308	0.2345	0.1638
	1-2	0.5071	0.5035	0.3718	0.3570	0.2935	0.2649	0.2798	0.2444
	1-2-3	0.5225	0.5192	0.3947	0.3762	0.3386	0.3307	0.3024	0.2927
DEP	1	0.3448	0.3216	0.2452	0.1760	0.2208	0.1383	0.2104	0.1335
	1-2	0.4338	0.4237	0.3042	0.2536	0.2771	0.2313	0.2341	0.1972
	1-2-3	0.4690	0.4677	0.3723	0.3681	0.3002	0.2653	0.2440	0.1781
LEMMA	1	0.5379	0.5390	0.4516	0.4475	0.3518	0.3446	0.3077	0.2764
	1-2	0.5433	0.5426	0.4860	0.4843	0.3675	0.3609	0.3102	0.3018
WORD	1	0.4998	0.5014	0.4085	0.4068	0.3011	0.2568	0.2697	0.2543
	1-2	0.5093	0.5080	0.4236	0.4178	0.3410	0.3344	0.3027	0.2893
NUM	1	0.2595	0.2423	0.2634	0.2383	0.2661	0.2353	0.2106	0.1918

Table E.11: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
CHAR	2	0.0865	0.0131	-0.0061	0.0224	-0.0068	0.0106
	2-3	0.0681	0.0186	0.0013	–	-0.0044	-0.0007
	2-3-4	0.0614	0.0082	0.0056	0.0025	0.0104	0.0068
	2-3-4-5	0.0544	0.0085	0.0076	0.0093	0.0013	-0.0022
ASIS	2	0.0697	-0.0046	0.0127	0.0164	-0.0052	-0.0028
	2-3	0.0731	0.0092	-0.0003	0.0086	0.0048	-0.0017
	2-3-4	0.0588	0.0016	0.0051	0.0058	0.0031	0.0043
	2-3-4-5	0.0546	0.0129	0.0092	0.0109	0.0028	0.0018
POS	1	0.3351	0.3351	0.3225	0.3225	-0.0642	-0.0642
	1-2	0.0164	0.0153	0.0506	0.0512	0.0775	0.0510
	1-2-3	0.0483	0.0510	0.0544	0.0511	-0.0014	0.0060
TAG	1	-0.0185	0.0673	0.0453	0.0453	0.0256	0.0256
	1-2	0.0452	0.0696	0.0474	0.0469	0.0464	0.0452
	1-2-3	0.0400	0.0834	0.0099	0.0097	0.0017	0.0006
DEP	1	0.1238	0.0699	0.0320	0.0320	-0.0798	-0.0798
	1-2	0.1172	0.0457	-0.0097	-0.0097	0.0150	0.0150
	1-2-3	0.0376	0.0561	0.0511	0.0444	0.0211	0.0173
LEMMA	1	-0.0878	0.0197	0.0073	0.0009	-0.0049	0.0194
	1-2	-0.0661	0.0356	0.0217	-0.0039	-0.0011	0.0040
WORD	1	-0.1184	0.0144	-0.0030	0.0417	-0.0089	-0.0003
	1-2	-0.0686	0.0209	0.0142	0.0203	-0.0035	0.0018
NUM	1	0.2100	0.2100	0.0900	0.0900	0.2667	0.2667

Minimum of Characters: 500

Table E.12: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.6367	0.6322	0.5043	0.3481	0.5095	0.3680	0.5582	0.5403
	2-3	0.6531	0.6477	0.5202	0.3788	0.5278	0.4365	0.5518	0.5173
	2-3-4	0.6672	0.6624	0.6266	0.6056	0.5609	0.5462	0.5636	0.5432
	2-3-4-5	0.6820	0.6789	0.6404	0.6402	0.5584	0.5385	0.5452	0.4858
CHAR	2	0.8080	0.8080	0.7590	0.7580	0.6901	0.6882	0.6720	0.6719
	2-3	0.8694	0.8690	0.8164	0.8151	0.7641	0.7638	0.7265	0.7260
	2-3-4	0.8713	0.8702	0.8413	0.8411	0.7751	0.7746	0.7369	0.7362
	2-3-4-5	0.8848	0.8847	0.8122	0.8094	0.7759	0.7756	0.7412	0.7411
ASIS	2	0.8366	0.8357	0.7996	0.7995	0.7186	0.7184	0.6925	0.6916
	2-3	0.8871	0.8868	0.8444	0.8436	0.7753	0.7750	0.7354	0.7341
	2-3-4	0.9006	0.9004	0.8664	0.8662	0.7886	0.7885	0.7494	0.7493
	2-3-4-5	0.9042	0.9040	0.8197	0.8158	0.7882	0.7882	0.7493	0.7488
POS	1	0.6032	0.5907	0.5864	0.5847	0.5776	0.5775	0.5730	0.5728
	1-2	0.6413	0.6303	0.6404	0.6403	0.6062	0.6061	0.5988	0.5985
	1-2-3	0.6597	0.6583	0.6587	0.6579	0.6189	0.6174	0.6128	0.6127
TAG	1	0.6410	0.6402	0.5930	0.5700	0.5893	0.5892	0.5772	0.5771
	1-2	0.6990	0.6982	0.6710	0.6704	0.6274	0.6265	0.6140	0.6140
	1-2-3	0.6987	0.6980	0.6830	0.6828	0.6299	0.6230	0.6279	0.6278
DEP	1	0.6219	0.6195	0.6017	0.5997	0.5691	0.5652	0.5678	0.5655
	1-2	0.6590	0.6590	0.6322	0.6277	0.5976	0.5963	0.5935	0.5926
	1-2-3	0.6705	0.6704	0.6566	0.6550	0.6053	0.6008	0.6066	0.6060
LEMMA	1	0.7788	0.7786	0.7680	0.7679	0.7096	0.7096	0.6876	0.6869
	1-2	0.8008	0.8007	0.7816	0.7816	0.7145	0.7144	0.6935	0.6925
WORD	1	0.7591	0.7588	0.7434	0.7408	0.6942	0.6940	0.6785	0.6782
	1-2	0.7653	0.7653	0.7535	0.7535	0.6983	0.6981	0.6837	0.6836
NUM	1	0.5950	0.5912	0.5635	0.5635	0.5542	0.5229	0.5332	0.4791

Table E.13: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
CHAR	2	0.2757	0.0476	0.0903	0.1012	0.0932	0.0932
	2-3	0.1739	0.0056	0.0277	0.0341	0.0251	0.0317
	2-3-4	0.1328	0.0002	0.0174	0.0285	0.0364	0.0302
	2-3-4-5	0.0999	-0.0035	0.0131	0.0221	0.0215	0.0244
ASIS	2	0.2720	0.0435	0.0573	0.0488	0.0588	0.0656
	2-3	0.1492	0.0154	0.0155	0.0369	0.0213	0.0390
	2-3-4	0.1229	-0.0056	0.0114	0.0142	0.0150	0.0139
	2-3-4-5	0.1025	0.0215	0.0207	0.0084	0.0153	0.0196
POS	1	0.2456	0.2456	0.3123	0.3123	0.1982	0.1982
	1-2	0.0272	0.0314	-0.1515	-0.1286	0.2639	0.2440
	1-2-3	-0.0449	-0.0707	-0.0222	-0.0120	0.0582	0.0586
TAG	1	-0.0086	-0.0086	-0.1640	-0.1460	0.0043	0.0043
	1-2	0.1546	0.1327	0.0233	-0.0280	0.1140	0.1091
	1-2-3	0.0514	0.0577	0.0017	-0.0141	0.1243	0.1212
DEP	1	0.0103	-0.1889	-0.0425	-0.0425	0.0244	0.0244
	1-2	0.1047	0.0867	-0.1028	-0.1028	-0.0255	-0.0422
	1-2-3	-0.0349	-0.0498	0.0236	0.0366	0.0737	0.0692
LEMMA	1	-0.0374	-0.0415	0.0221	0.0459	0.0406	-0.0087
	1-2	-0.0143	0.0461	0.0099	-0.0155	0.0467	-0.0293
WORD	1	-0.0304	0.0282	0.0136	-0.0115	0.0480	0.0523
	1-2	-0.0003	0.0252	0.0560	0.0179	0.0304	-0.0080
NUM	1	0.2500	0.2500	0.0833	0.0833	0.0667	0.0667

Table E.14: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST	2	0.4831	0.4797	0.2924	0.2249	0.3224	0.2808	0.2819	0.2196
	2-3	0.4969	0.4809	0.3987	0.3711	0.3089	0.2633	0.3029	0.2736
	2-3-4	0.5338	0.5294	0.4561	0.4482	0.2905	0.2454	0.3208	0.2921
	2-3-4-5	0.5427	0.5384	0.4385	0.4088	0.3021	0.2401	0.3217	0.2823
CHAR	2	0.7240	0.7257	0.6254	0.6250	0.4829	0.4761	0.4143	0.4033
	2-3	0.8023	0.8024	0.7320	0.7319	0.5762	0.5696	0.5104	0.5070
	2-3-4	0.8268	0.8274	0.7575	0.7579	0.6009	0.5938	0.5368	0.5346
	2-3-4-5	0.8290	0.8291	0.7494	0.7492	0.6136	0.6114	0.5439	0.5407
ASIS	2	0.7801	0.7811	0.6779	0.6776	0.5188	0.5142	0.4515	0.4426
	2-3	0.8540	0.8546	0.7651	0.7646	0.6086	0.6048	0.5371	0.5346
	2-3-4	0.8682	0.8687	0.7843	0.7836	0.6253	0.6189	0.5559	0.5512
	2-3-4-5	0.8544	0.8542	0.7893	0.7890	0.6236	0.6174	0.5572	0.5523
POS	1	0.3740	0.3529	0.2990	0.2723	0.2140	0.0964	0.2056	0.1010
	1-2	0.4875	0.4794	0.3689	0.3512	0.2864	0.2269	0.2678	0.2283
	1-2-3	0.5321	0.5248	0.4249	0.4124	0.3220	0.2767	0.2844	0.2491
TAG	1	0.4083	0.3934	0.3316	0.3244	0.2301	0.1338	0.2291	0.1690
	1-2	0.5476	0.5432	0.4492	0.4479	0.3318	0.2755	0.2998	0.2670
	1-2-3	0.5877	0.5878	0.4856	0.4859	0.3875	0.3811	0.3402	0.3110
DEP	1	0.3704	0.3560	0.2994	0.2418	0.2089	0.1000	0.2504	0.2387
	1-2	0.5089	0.5057	0.3869	0.3820	0.3140	0.2523	0.2762	0.2358
	1-2-3	0.5370	0.5369	0.4142	0.4029	0.3468	0.3292	0.3108	0.2873
LEMMA	1	0.6679	0.6668	0.5773	0.5763	0.4239	0.4091	0.3651	0.3339
	1-2	0.6923	0.6920	0.5924	0.5881	0.4493	0.4424	0.3916	0.3781
WORD	1	0.6282	0.6282	0.5433	0.5416	0.3993	0.3867	0.3523	0.3439
	1-2	0.6438	0.6437	0.5524	0.5482	0.4168	0.4053	0.3551	0.3294
NUM	1	0.3099	0.2995	0.2714	0.2542	0.2807	0.2386	0.2440	0.2194

Table E.15: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a feature-wise model on the individual feature types

	Target	Age					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
CHAR	2	0.2159	0.0025	0.0197	0.0100	0.0058	-0.0009
	2-3	0.1388	0.0013	0.0014	0.0022	-0.0014	0.0042
	2-3-4	0.1140	0.0123	0.0124	0.0118	0.0123	0.0141
	2-3-4-5	0.0916	0.0113	0.0075	0.0069	0.0068	0.0055
ASIS	2	0.2031	0.0121	0.0049	0.0141	-0.0024	0.0076
	2-3	0.1386	0.0059	0.0027	-0.0003	0.0025	0.0066
	2-3-4	0.1080	0.0070	0.0032	0.0008	0.0069	0.0038
	2-3-4-5	0.0925	0.0066	0.0023	0.0045	0.0011	0.0056
POS	1	0.1807	0.1807	0.1046	0.1046	0.1923	0.1923
	1-2	0.0762	0.0317	-0.0211	-0.0158	0.0116	0.0116
	1-2-3	0.0611	0.0675	0.0236	0.0195	0.0455	0.0432
TAG	1	0.0886	0.0509	0.0711	0.0010	-0.0680	-0.0357
	1-2	0.0178	-0.0089	-0.0109	-0.0351	0.0441	0.0250
	1-2-3	0.0825	0.0528	0.0386	0.0268	0.0474	0.0551
DEP	1	0.0971	0.1030	0.0466	0.0466	0.0411	0.0411
	1-2	0.0559	0.0636	0.0155	0.0068	-0.0100	-0.0088
	1-2-3	0.0751	0.0724	-0.0095	-0.0003	-0.0020	0.0079
LEMMA	1	-0.0219	0.0181	0.0128	-0.0047	0.0078	0.0045
	1-2	0.0010	-0.0082	0.0085	0.0027	-0.0058	0.0056
WORD	1	-0.0585	0.0064	0.0428	0.0232	0.0151	0.0098
	1-2	-0.0112	-0.0047	0.0013	0.0081	0.0118	0.0092
NUM	1	0.4267	0.4267	-0.0933	-0.0933	0.3033	0.3033

DIST, CHAR, ASIS, WORD, and LEMMA

Minimum of Characters: 100 & Cumulated

Table E.16: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7343	0.7338	0.6967	0.6947	0.6284	0.6276	0.6089	0.5966
	2-3	0.7746	0.7738	0.7331	0.7330	0.6647	0.6645	0.6387	0.6301
	2-3-4	0.7882	0.7881	0.7379	0.7370	0.6712	0.6711	0.6412	0.6312
	2-3-4-5	0.7836	0.7829	0.7354	0.7342	0.6710	0.6710	0.6409	0.6305
DIST_CHAR_ASIS	2	0.7905	0.7904	0.7453	0.7453	0.6775	0.6775	0.6452	0.6340
	2-3	0.7973	0.7972	0.7537	0.7534	0.6835	0.6833	0.6645	0.6645
	2-3-4	0.7913	0.7908	0.7478	0.7459	0.6849	0.6848	0.6660	0.6660
	2-3-4-5	0.7980	0.7979	0.7461	0.7437	0.6844	0.6844	0.6658	0.6658
DIST_CHAR_ASIS_LEMMA	1	0.8014	0.8013	0.7550	0.7549	0.6868	0.6861	0.6688	0.6688
	1-2	0.7999	0.7998	0.7541	0.7534	0.6882	0.6882	0.6689	0.6689
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8010	0.8007	0.7550	0.7548	0.6828	0.6801	0.6700	0.6694
	1-2	0.8045	0.8044	0.7567	0.7566	0.6832	0.6803	0.6707	0.6702

Table E.17: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
DIST_CHAR	2	-0.0481	-0.0285	-0.0020	-0.0628	0.0686	0.0497
	2-3	-0.0355	0.0028	0.0120	0.0464	0.0665	0.0553
	2-3-4	-0.0022	0.0259	0.0037	0.0052	0.0198	0.0238
	2-3-4-5	-0.0103	0.0157	0.0228	0.0138	0.0165	0.0021
DIST_CHAR_ASIS	2	-0.0215	0.0133	-0.0312	0.0060	0.0065	0.0232
	2-3	-0.0055	-0.0018	0.0044	0.0263	-0.0050	0.0227
	2-3-4	-0.0265	-0.0274	0.0009	-0.0053	0.0245	0.0168
	2-3-4-5	-0.0284	-0.0212	-0.0118	-0.0198	-0.0143	-0.0140
DIST_CHAR_ASIS_LEMMA	1	-0.0142	-0.0172	0.0200	0.0075	0.0022	0.0267
	1-2	0.0151	-0.0069	0.0094	-0.0110	0.0216	0.0211
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0075	0.0216	0.0092	-0.0109	-0.0022	0.0149
	1-2	0.0111	0.0057	0.0093	-0.0034	0.0189	0.0100

Table E.18: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.5983	0.5994	0.4887	0.4861	0.3739	0.3555	0.3088	0.2673
	2-3	0.6907	0.6908	0.5595	0.5583	0.4424	0.4382	0.3888	0.3706
	2-3-4	0.7031	0.7037	0.5749	0.5714	0.4523	0.4435	0.4023	0.3817
	2-3-4-5	0.6804	0.6827	0.5378	0.5386	0.4624	0.4589	0.4017	0.3870
DIST_CHAR_ASIS	2	0.7161	0.7161	0.5828	0.5810	0.4677	0.4662	0.4223	0.4062
	2-3	0.7275	0.7276	0.5944	0.5926	0.4759	0.4726	0.4336	0.4241
	2-3-4	0.7270	0.7267	0.6014	0.6002	0.4744	0.4672	0.4309	0.4160
	2-3-4-5	0.7292	0.7295	0.5952	0.5939	0.4788	0.4738	0.4320	0.4203
DIST_CHAR_ASIS_LEMMA	1	0.7056	0.7042	0.5915	0.5902	0.4850	0.4799	0.4372	0.4295
	1-2	0.7213	0.7220	0.5950	0.5942	0.4838	0.4801	0.4377	0.4299
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7261	0.7265	0.5994	0.5989	0.4845	0.4818	0.4329	0.4201
	1-2	0.7218	0.7221	0.6055	0.6053	0.4842	0.4813	0.4356	0.4243

Table E.19: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
DIST_CHAR	2	0.0013	-0.0092	0.0077	0.0500	0.0047	0.0328
	2-3	0.0064	0.0009	-0.0084	-0.0053	0.0196	0.0099
	2-3-4	-0.0009	0.0049	0.0101	0.0123	-0.0004	-0.0037
	2-3-4-5	0.0157	0.0139	0.0107	-0.0099	0.0016	0.0008
DIST_CHAR_ASIS	2	0.0108	0.0058	-0.0019	0.0102	0.0090	0.0187
	2-3	0.0089	0.0029	0.0068	0.0036	0.0053	0.0122
	2-3-4	0.0119	0.0045	0.0045	0.0047	0.0087	0.0154
	2-3-4-5	0.0074	0.0087	-0.0049	-0.0032	0.0018	0.0034
DIST_CHAR_ASIS_LEMMA	1	0.0026	0.0128	0.0053	0.0030	0.0128	0.0147
	1-2	0.0072	0.0095	-0.0051	-0.0010	0.0127	0.0073
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0039	-0.0016	0.0130	0.0050	0.0054	0.0082
	1-2	0.0071	0.0105	-0.0009	-0.0024	0.0037	0.0097

Minimum of Characters: 250 & Cumulated

Table E.20: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7844	0.7841	0.7495	0.7493	0.6736	0.6736	0.6514	0.6511
	2-3	0.8249	0.8247	0.7771	0.7738	0.7219	0.7218	0.6916	0.6916
	2-3-4	0.8310	0.8309	0.7884	0.7858	0.7306	0.7305	0.7001	0.7001
	2-3-4-5	0.8316	0.8315	0.7943	0.7927	0.7283	0.7279	0.6995	0.6995
DIST_CHAR_ASIS	2	0.8434	0.8434	0.8019	0.8002	0.7419	0.7418	0.7084	0.7084
	2-3	0.8423	0.8418	0.8120	0.8104	0.7475	0.7473	0.7166	0.7165
	2-3-4	0.8499	0.8499	0.8260	0.8260	0.7497	0.7495	0.7176	0.7172
	2-3-4-5	0.8479	0.8479	0.8243	0.8243	0.7485	0.7484	0.7173	0.7168
DIST_CHAR_ASIS_LEMMA	1	0.8477	0.8477	0.8238	0.8237	0.7350	0.7298	0.7221	0.7221
	1-2	0.8508	0.8508	0.8174	0.8165	0.7362	0.7311	0.7228	0.7228
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8470	0.8470	0.8238	0.8237	0.7552	0.7552	0.7220	0.7218
	1-2	0.8472	0.8471	0.8250	0.8250	0.7553	0.7553	0.7222	0.7218

Table E.21: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
DIST_CHAR	2	0.0017	0.0251	0.0348	0.0854	0.0363	0.0600
	2-3	-0.0013	0.0141	-0.0077	0.0357	0.0245	0.0117
	2-3-4	0.0015	0.0345	0.0257	0.0152	0.0389	0.0345
	2-3-4-5	-0.0060	0.0096	0.0168	0.0283	0.0403	0.0370
DIST_CHAR_ASIS	2	0.0181	0.0458	0.0196	-0.0173	-0.0216	-0.0147
	2-3	-0.0017	0.0099	0.0257	0.0034	0.0041	-0.0010
	2-3-4	0.0151	-0.0199	0.0072	-0.0106	0.0263	0.0332
	2-3-4-5	0.0060	0.0030	0.0127	-0.0121	0.0074	-0.0042
DIST_CHAR_ASIS_LEMMA	1	0.0061	0.0041	0.0156	0.0043	0.0136	0.0103
	1-2	0.0062	-0.0082	0.0058	-0.0069	0.0152	0.0150
DIST_CHAR_ASIS_LEMMA_WORD	1	-0.0095	0.0059	0.0053	0.0094	0.0108	0.0203
	1-2	-0.0079	-0.0089	0.0166	0.0134	0.0077	0.0049

Table E.22: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.6772	0.6770	0.5772	0.5772	0.4543	0.4439	0.3928	0.3686
	2-3	0.7627	0.7629	0.6684	0.6689	0.5239	0.5115	0.4643	0.4581
	2-3-4	0.7534	0.7555	0.6707	0.6732	0.5471	0.5440	0.4860	0.4795
	2-3-4-5	0.7380	0.7413	0.6847	0.6848	0.5454	0.5361	0.4927	0.4875
DIST_CHAR_ASIS	2	0.7749	0.7751	0.6993	0.6996	0.5607	0.5567	0.4827	0.4712
	2-3	0.7918	0.7920	0.7083	0.7067	0.5742	0.5705	0.5002	0.4942
	2-3-4	0.7981	0.7996	0.7140	0.7131	0.5773	0.5736	0.4982	0.4902
	2-3-4-5	0.7928	0.7941	0.7059	0.7071	0.5862	0.5845	0.5049	0.5020
DIST_CHAR_ASIS_LEMMA	1	0.7737	0.7743	0.7069	0.7054	0.5857	0.5829	0.5066	0.5014
	1-2	0.7867	0.7863	0.7075	0.7079	0.5782	0.5728	0.5060	0.4989
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7774	0.7778	0.7030	0.7037	0.5803	0.5757	0.5056	0.4972
	1-2	0.7911	0.7916	0.7070	0.7074	0.5862	0.5820	0.5082	0.5022

Table E.23: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
DIST_CHAR	2	0.0620	0.0267	0.0122	0.0015	0.0181	0.0090
	2-3	0.0686	0.0548	0.0189	0.0225	0.0152	0.0079
	2-3-4	0.0693	0.0148	0.0045	0.0058	0.0072	0.0068
	2-3-4-5	0.0496	0.0082	0.0093	0.0005	-0.0054	-0.0113
DIST_CHAR_ASIS	2	0.0390	0.0173	0.0187	0.0134	0.0050	0.0077
	2-3	0.0346	0.0165	0.0025	0.0088	0.0027	0.0066
	2-3-4	0.0333	0.0131	0.0004	0.0048	0.0168	0.0067
	2-3-4-5	0.0450	0.0078	-0.0004	0.0097	0.0206	-0.0040
DIST_CHAR_ASIS_LEMMA	1	0.0474	0.0163	0.0010	0.0054	0.0083	0.0038
	1-2	0.0445	0.0153	0.0165	0.0143	0.0040	-0.0025
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0495	0.0140	0.0119	0.0137	0.0022	0.0153
	1-2	0.0449	0.0180	0.0132	0.0067	0.0151	0.0067

Minimum of Characters: 500 & Cumulated

Table E.24: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.8231	0.8230	0.7868	0.7868	0.7062	0.7058	0.6795	0.6769
	2-3	0.8802	0.8802	0.8396	0.8392	0.7698	0.7696	0.7313	0.7311
	2-3-4	0.8727	0.8712	0.8498	0.8494	0.7852	0.7852	0.7407	0.7392
	2-3-4-5	0.8953	0.8951	0.8583	0.8583	0.7679	0.7650	0.7433	0.7429
DIST_CHAR_ASIS	2	0.8887	0.8882	0.8623	0.8623	0.7521	0.7441	0.7511	0.7494
	2-3	0.8861	0.8850	0.8725	0.8725	0.8036	0.8036	0.7612	0.7602
	2-3-4	0.8779	0.8764	0.8734	0.8733	0.7700	0.7641	0.7676	0.7675
	2-3-4-5	0.8947	0.8942	0.8730	0.8729	0.7688	0.7626	0.7674	0.7673
DIST_CHAR_ASIS_LEMMA	1	0.8930	0.8925	0.8697	0.8697	0.7805	0.7765	0.7707	0.7706
	1-2	0.8782	0.8779	0.8744	0.8744	0.7785	0.7740	0.7708	0.7703
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8799	0.8797	0.8712	0.8712	0.8037	0.8034	0.7698	0.7688
	1-2	0.8943	0.8943	0.8704	0.8702	0.7754	0.7702	0.7727	0.7723

Table E.25: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
DIST_CHAR	2	0.0911	0.0265	0.0598	0.0466	0.0998	0.0889
	2-3	0.0903	-0.0365	0.0331	0.0278	0.0178	0.0419
	2-3-4	0.0990	0.0314	0.0445	0.0261	0.0454	0.0362
	2-3-4-5	0.0829	-0.0121	0.0455	0.0365	0.0461	0.0232
DIST_CHAR_ASIS	2	0.0815	-0.0217	0.0306	0.0077	0.0118	0.0141
	2-3	0.0931	0.0012	0.0462	0.0171	0.0295	0.0138
	2-3-4	0.0789	-0.0137	0.0380	0.0166	0.0026	0.0088
	2-3-4-5	0.0829	0.0014	0.0459	0.0050	0.0281	0.0243
DIST_CHAR_ASIS_LEMMA	1	0.0935	0.0170	0.0201	0.0087	0.0152	0.0141
	1-2	0.0829	0.0101	0.0428	0.0172	0.0084	0.0070
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0632	0.0040	0.0388	0.0146	-0.0029	0.0213
	1-2	0.0526	0.0023	0.0268	0.0127	0.0094	0.0274

Table E.26: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7262	0.7271	0.6350	0.6349	0.5133	0.5123	0.4465	0.4434
	2-3	0.8192	0.8214	0.7342	0.7331	0.5938	0.5898	0.5199	0.5177
	2-3-4	0.8321	0.8343	0.7623	0.7627	0.6151	0.6114	0.5437	0.5390
	2-3-4-5	0.8232	0.8250	0.7590	0.7595	0.6186	0.6140	0.5451	0.5397
DIST_CHAR_ASIS	2	0.8508	0.8518	0.7725	0.7727	0.6203	0.6150	0.5566	0.5515
	2-3	0.8727	0.8726	0.7954	0.7954	0.6640	0.6629	0.5724	0.5687
	2-3-4	0.8753	0.8753	0.8017	0.8014	0.6618	0.6610	0.5643	0.5553
	2-3-4-5	0.8673	0.8671	0.8007	0.8006	0.6421	0.6376	0.5764	0.5731
DIST_CHAR_ASIS_LEMMA	1	0.8611	0.8608	0.7971	0.7966	0.6518	0.6478	0.5773	0.5717
	1-2	0.8620	0.8615	0.7961	0.7963	0.6655	0.6629	0.5643	0.5544
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8517	0.8515	0.7885	0.7883	0.6585	0.6558	0.5771	0.5736
	1-2	0.8290	0.8293	0.7736	0.7727	0.6653	0.6652	0.5798	0.5758

Table E.27: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
DIST_CHAR	2	0.1366	0.0099	0.0244	0.0326	0.0048	0.0061
	2-3	0.1364	0.0224	0.0286	0.0257	0.0117	0.0144
	2-3-4	0.1078	0.0159	0.0183	0.0196	0.0110	-0.0038
	2-3-4-5	0.0934	0.0227	0.0179	0.0167	0.0084	0.0023
DIST_CHAR_ASIS	2	0.0901	-0.0056	0.0189	0.0107	0.0071	0.0017
	2-3	0.0815	0.0144	0.0157	0.0055	0.0102	0.0100
	2-3-4	0.0848	0.0158	0.0124	0.0120	0.0053	0.0009
	2-3-4-5	0.0831	0.0126	0.0198	0.0020	0.0067	0.0095
DIST_CHAR_ASIS_LEMMA	1	0.0753	0.0081	0.0143	0.0131	0.0077	0.0040
	1-2	0.0736	0.0087	0.0197	0.0150	0.0028	0.0084
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0680	0.0152	0.0044	0.0052	0.0056	0.0053
	1-2	0.0682	0.0210	0.0116	0.0111	0.0095	0.0098

Minimum of Characters: 100 & Stacked

Table E.28: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7223	0.7220	0.6879	0.6864	0.6232	0.6230	0.6142	0.6136
	2-3	0.7658	0.7658	0.7159	0.7139	0.6594	0.6594	0.6426	0.6424
	2-3-4	0.7776	0.7776	0.7259	0.7253	0.6655	0.6655	0.6478	0.6475
	2-3-4-5	0.7732	0.7731	0.7249	0.7243	0.6638	0.6637	0.6477	0.6474
DIST_CHAR_ASIS	2	0.7761	0.7760	0.7217	0.7200	0.6657	0.6656	0.6487	0.6484
	2-3	0.7881	0.7880	0.7314	0.7300	0.6694	0.6693	0.6518	0.6516
	2-3-4	0.7914	0.7914	0.7361	0.7348	0.6723	0.6723	0.6543	0.6540
	2-3-4-5	0.7890	0.7889	0.7338	0.7324	0.6708	0.6707	0.6540	0.6539
DIST_CHAR_ASIS_LEMMA	1	0.7898	0.7897	0.7342	0.7329	0.6707	0.6707	0.6537	0.6535
	1-2	0.7913	0.7912	0.7349	0.7336	0.6702	0.6700	0.6537	0.6535
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7909	0.7909	0.7360	0.7347	0.6726	0.6726	0.6550	0.6548
	1-2	0.7906	0.7903	0.7357	0.7345	0.6722	0.6722	0.6549	0.6546

Table E.29: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	-0.0387	-0.0005	0.0055	-0.0134	0.0438	0.0420
	2-3	-0.0308	-0.0107	-0.0109	-0.0139	0.0442	0.0409
	2-3-4	-0.0293	-0.0093	-0.0037	-0.0112	0.0369	0.0404
	2-3-4-5	-0.0168	0.0059	-0.0041	-0.0094	0.0306	0.0362
DIST_CHAR_ASIS	2	-0.0102	0.0104	-0.0034	-0.0049	0.0243	0.0322
	2-3	-0.0078	0.0036	0.0263	-0.0042	-0.0069	0.0140
	2-3-4	-0.0236	-0.0164	-0.0019	-0.0071	0.0207	0.0259
	2-3-4-5	-0.0155	-0.0105	-0.0049	-0.0039	0.0203	0.0195
DIST_CHAR_ASIS_LEMMA	1	-0.0120	-0.0057	-0.0043	0.0013	0.0177	0.0192
	1-2	-0.0053	0.0045	-0.0064	-0.0022	0.0154	0.0157
DIST_CHAR_ASIS_LEMMA_WORD	1	-0.0031	0.0093	-0.0053	0.0002	0.0149	0.0158
	1-2	-0.0069	-0.0138	-0.0088	-0.0012	0.0225	0.0174

Table E.30: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.5814	0.5806	0.4549	0.4544	0.2092	0.0892	0.2029	0.0707
	2-3	0.6560	0.6565	0.5434	0.5426	0.4313	0.4281	0.2607	0.1626
	2-3-4	0.6726	0.6728	0.5630	0.5621	0.4346	0.4147	0.3547	0.2988
	2-3-4-5	0.6769	0.6773	0.5647	0.5638	0.4531	0.4464	0.3450	0.2828
DIST_CHAR_ASIS	2	0.6955	0.6954	0.5636	0.5627	0.4507	0.4473	0.3738	0.3267
	2-3	0.7126	0.7128	0.5758	0.5750	0.4604	0.4584	0.3820	0.3378
	2-3-4	0.7164	0.7163	0.5829	0.5822	0.4703	0.4649	0.3707	0.3001
	2-3-4-5	0.7156	0.7156	0.5835	0.5828	0.4695	0.4638	0.3786	0.3584
DIST_CHAR_ASIS_LEMMA	1	0.7158	0.7158	0.5845	0.5839	0.4703	0.4644	0.3673	0.3416
	1-2	0.7178	0.7177	0.5836	0.5829	0.4704	0.4645	0.3679	0.3422
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7168	0.7167	0.5840	0.5832	0.4706	0.4645	0.3827	0.3481
	1-2	0.7173	0.7172	0.5837	0.5830	0.4706	0.4646	0.3810	0.3449

Table E.31: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	-0.0261	-0.0180	-0.0041	0.0083	0.0156	0.0209
	2-3	-0.0058	-0.0031	-0.0170	-0.0032	0.0285	0.0297
	2-3-4	-0.0065	0.0028	0.0036	0.0192	-0.0086	-0.0014
	2-3-4-5	-0.0145	-0.0003	-0.0099	0.0049	0.0225	0.0185
DIST_CHAR_ASIS	2	0.0049	-0.0039	0.0051	0.0080	0.0112	0.0043
	2-3	-0.0002	-0.0051	-0.0048	0.0081	0.0121	0.0122
	2-3-4	0.0066	0.0059	-0.0038	0.0006	0.0089	0.0184
	2-3-4-5	0.0106	0.0062	-0.0028	0.0105	0.0121	0.0134
DIST_CHAR_ASIS_LEMMA	1	0.0093	0.0081	0.0066	0.0132	0.0015	0.0014
	1-2	0.0021	0.0123	0.0027	0.0058	0.0019	0.0066
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0032	-0.0001	0.0112	0.0060	0.0075	0.0043
	1-2	-0.0002	0.0096	0.0055	0.0137	0.0080	0.0116

Minimum of Characters: 250 & Stacked

Table E.32: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7737	0.7735	0.7414	0.7414	0.6647	0.6642	0.6455	0.6436
	2-3	0.8138	0.8130	0.7813	0.7813	0.7156	0.7156	0.6858	0.6849
	2-3-4	0.8218	0.8217	0.7912	0.7911	0.7261	0.7261	0.6933	0.6926
	2-3-4-5	0.8230	0.8230	0.7927	0.7926	0.7266	0.7266	0.6936	0.6928
DIST_CHAR_ASIS	2	0.8287	0.8287	0.7983	0.7983	0.7282	0.7280	0.6947	0.6940
	2-3	0.8397	0.8396	0.8076	0.8076	0.7323	0.7323	0.6996	0.6990
	2-3-4	0.8392	0.8391	0.8108	0.8106	0.7353	0.7353	0.7051	0.7050
	2-3-4-5	0.8359	0.8357	0.8114	0.8113	0.7345	0.7345	0.7046	0.7045
DIST_CHAR_ASIS_LEMMA	1	0.8428	0.8427	0.8090	0.8088	0.7338	0.7338	0.7041	0.7041
	1-2	0.8381	0.8379	0.8109	0.8108	0.7345	0.7345	0.7044	0.7043
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8387	0.8385	0.8118	0.8117	0.7370	0.7369	0.7050	0.7044
	1-2	0.8426	0.8425	0.8126	0.8125	0.7365	0.7364	0.7054	0.7054

Table E.33: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.0315	0.0538	0.0180	0.0214	0.0739	0.0638
	2-3	0.0250	0.0498	0.0184	0.0121	0.0694	0.0464
	2-3-4	0.0295	0.0412	0.0155	0.0154	0.0480	0.0440
	2-3-4-5	0.0248	0.0451	0.0103	0.0112	0.0500	0.0452
DIST_CHAR_ASIS	2	0.0336	0.0458	0.0110	0.0130	0.0486	0.0443
	2-3	0.0315	0.0388	0.0124	0.0391	0.0262	0.0248
	2-3-4	0.0327	0.0429	0.0126	0.0066	0.0414	0.0400
	2-3-4-5	0.0557	-0.0074	0.0112	0.0075	0.0403	0.0336
DIST_CHAR_ASIS_LEMMA	1	0.0366	-0.0120	0.0111	0.0280	0.0339	0.0347
	1-2	0.0447	-0.0151	0.0100	0.0073	0.0339	0.0312
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0380	-0.0132	0.0121	0.0084	0.0332	0.0308
	1-2	0.0207	0.0447	0.0108	0.0081	0.0305	0.0263

Table E.34: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.6667	0.6672	0.5623	0.5621	0.4216	0.4205	0.3466	0.3196
	2-3	0.7480	0.7484	0.6506	0.6498	0.5183	0.5180	0.4535	0.4513
	2-3-4	0.7588	0.7592	0.6741	0.6735	0.5337	0.5308	0.4774	0.4757
	2-3-4-5	0.7627	0.7630	0.6722	0.6718	0.5325	0.5292	0.4829	0.4793
DIST_CHAR_ASIS	2	0.7759	0.7763	0.6793	0.6788	0.5455	0.5419	0.4849	0.4810
	2-3	0.7940	0.7943	0.6919	0.6914	0.5571	0.5567	0.4930	0.4892
	2-3-4	0.7911	0.7915	0.6996	0.6991	0.5553	0.5529	0.4994	0.4962
	2-3-4-5	0.7933	0.7933	0.6966	0.6962	0.5553	0.5530	0.5000	0.4972
DIST_CHAR_ASIS_LEMMA	1	0.7947	0.7948	0.6989	0.6986	0.5562	0.5540	0.5001	0.4976
	1-2	0.7959	0.7961	0.6956	0.6949	0.5596	0.5559	0.5002	0.4977
DIST_CHAR_ASIS_LEMMA_WORD	1	0.7962	0.7964	0.6963	0.6956	0.5641	0.5630	0.5006	0.4985
	1-2	0.7977	0.7979	0.6961	0.6954	0.5613	0.5579	0.4975	0.4967

Table E.35: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.0132	-0.0179	-0.0008	0.0021	0.0057	0.0131
	2-3	0.0076	-0.0155	0.0017	0.0131	0.0161	0.0122
	2-3-4	0.0248	0.0128	-0.0031	0.0107	0.0037	0.0109
	2-3-4-5	0.0154	—	0.0055	0.0092	0.0169	0.0218
DIST_CHAR_ASIS	2	0.0218	0.0024	-0.0060	-0.0045	0.0118	0.0127
	2-3	0.0311	-0.0075	0.0108	0.0040	0.0065	0.0082
	2-3-4	0.0289	-0.0034	0.0092	0.0006	-0.0009	-0.0029
	2-3-4-5	0.0335	0.0136	0.0001	0.0041	0.0002	—
DIST_CHAR_ASIS_LEMMA	1	0.0249	0.0066	0.0030	-0.0050	0.0019	0.0093
	1-2	0.0185	-0.0032	0.0022	-0.0045	0.0050	0.0150
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0150	0.0033	0.0011	0.0009	0.0057	0.0093
	1-2	0.0094	0.0105	0.0036	-0.0011	0.0085	0.0197

Minimum of Characters: 500 & Stacked

Table E.36: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.8192	0.8179	0.7780	0.7778	0.6948	0.6935	0.6765	0.6764
	2-3	0.8743	0.8740	0.8369	0.8369	0.7639	0.7636	0.7286	0.7285
	2-3-4	0.8874	0.8874	0.8485	0.8485	0.7772	0.7769	0.7400	0.7399
	2-3-4-5	0.8884	0.8883	0.8501	0.8500	0.7773	0.7769	0.7411	0.7410
DIST_CHAR_ASIS	2	0.8920	0.8919	0.8527	0.8527	0.7767	0.7761	0.7412	0.7411
	2-3	0.8983	0.8982	0.8619	0.8619	0.7844	0.7841	0.7467	0.7466
	2-3-4	0.8832	0.8830	0.8670	0.8670	0.7862	0.7855	0.7505	0.7504
	2-3-4-5	0.8789	0.8787	0.8664	0.8663	0.7856	0.7850	0.7515	0.7513
DIST_CHAR_ASIS_LEMMA	1	0.9015	0.9013	0.8665	0.8665	0.7852	0.7846	0.7522	0.7521
	1-2	0.8983	0.8979	0.8662	0.8661	0.7858	0.7852	0.7521	0.7520
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8992	0.8992	0.8666	0.8666	0.7863	0.7856	0.7524	0.7524
	1-2	0.8986	0.8983	0.8664	0.8663	0.7862	0.7855	0.7519	0.7519

Table E.37: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.1274	0.0535	0.0573	0.0718	0.0552	0.0749
	2-3	0.1182	0.0385	0.0420	0.0543	0.0388	0.0574
	2-3-4	0.1085	0.0315	0.0355	0.0490	0.0417	0.0531
	2-3-4-5	0.0951	0.0258	0.0311	0.0432	0.0336	0.0474
DIST_CHAR_ASIS	2	0.1148	0.0277	0.0340	0.0439	0.0364	0.0494
	2-3	0.1371	-0.0178	0.0286	0.0425	0.0308	0.0454
	2-3-4	0.1046	0.0192	0.0263	0.0358	0.0282	0.0381
	2-3-4-5	0.0987	0.0262	0.0281	0.0320	0.0272	0.0380
DIST_CHAR_ASIS_LEMMA	1	0.1079	-0.0122	0.0277	0.0331	0.0282	0.0344
	1-2	0.0976	-0.0065	0.0255	0.0252	0.0299	0.0284
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0886	-0.0027	0.0247	0.0228	0.0312	0.0300
	1-2	0.0831	-0.0038	0.0293	0.0243	0.0300	0.0239

Table E.38: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7413	0.7422	0.6347	0.6350	0.4899	0.4848	0.4263	0.4239
	2-3	0.8250	0.8252	0.7357	0.7356	0.5868	0.5840	0.5140	0.5114
	2-3-4	0.8415	0.8416	0.7623	0.7622	0.6156	0.6141	0.5433	0.5432
	2-3-4-5	0.8397	0.8397	0.7605	0.7603	0.6172	0.6161	0.5433	0.5416
DIST_CHAR_ASIS	2	0.8473	0.8472	0.7703	0.7701	0.6201	0.6187	0.5489	0.5469
	2-3	0.8651	0.8652	0.7795	0.7795	0.6280	0.6258	0.5550	0.5537
	2-3-4	0.8749	0.8748	0.7912	0.7911	0.6400	0.6396	0.5630	0.5630
	2-3-4-5	0.8700	0.8700	0.7878	0.7876	0.6372	0.6357	0.5623	0.5623
DIST_CHAR_ASIS_LEMMA	1	0.8700	0.8696	0.7843	0.7834	0.6392	0.6368	0.5623	0.5624
	1-2	0.8727	0.8723	0.7833	0.7824	0.6391	0.6366	0.5628	0.5628
DIST_CHAR_ASIS_LEMMA_WORD	1	0.8695	0.8691	0.7899	0.7897	0.6389	0.6365	0.5633	0.5634
	1-2	0.8722	0.8725	0.7820	0.7810	0.6382	0.6359	0.5629	0.5630

Table E.39: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the reduced feature set (DIST, CHAR, LEMMA, WORD)

	Target	Age					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.0733	0.0228	0.0064	0.0045	-0.0029	0.0157
	2-3	0.0645	-0.0005	0.0075	0.0223	-0.0016	0.0015
	2-3-4	0.0493	0.0004	0.0092	0.0073	0.0037	0.0017
	2-3-4-5	0.0451	0.0075	0.0236	0.0156	-0.0009	0.0045
DIST_CHAR_ASIS	2	0.0799	0.0056	0.0117	0.0167	0.0066	0.0003
	2-3	0.0807	0.0021	0.0149	0.0020	0.0013	0.0205
	2-3-4	0.0751	0.0083	-0.0057	-0.0054	0.0046	0.0045
	2-3-4-5	0.0688	-0.0053	0.0013	0.0040	0.0018	0.0051
DIST_CHAR_ASIS_LEMMA	1	0.0628	0.0107	0.0042	0.0054	0.0044	-0.0015
	1-2	0.0622	0.0067	-0.0011	0.0012	-0.0054	0.0015
DIST_CHAR_ASIS_LEMMA_WORD	1	0.0553	0.0101	0.0061	0.0045	-0.0038	0.0071
	1-2	0.0517	0.0057	0.0058	0.0032	-0.0036	0.0008

Full Set of feature types

Minimum of Characters: 100 & Cumulated

Table E.40: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7343	0.7338	0.6967	0.6947	0.6284	0.6276	0.6089	0.5966
	2-3	0.7746	0.7738	0.7331	0.7330	0.6647	0.6645	0.6387	0.6301
	2-3-4	0.7882	0.7881	0.7379	0.7370	0.6712	0.6711	0.6412	0.6312
	2-3-4-5	0.7836	0.7829	0.7354	0.7342	0.6710	0.6710	0.6409	0.6305
DIST_CHAR_ASIS	2	0.7905	0.7904	0.7453	0.7453	0.6775	0.6775	0.6452	0.6340
	2-3	0.7973	0.7972	0.7537	0.7534	0.6835	0.6833	0.6645	0.6645
	2-3-4	0.7913	0.7908	0.7478	0.7459	0.6849	0.6848	0.6660	0.6660
	2-3-4-5	0.7980	0.7979	0.7461	0.7437	0.6844	0.6844	0.6658	0.6658
DIST_CHAR_ASIS_POS	1	0.7994	0.7994	0.7519	0.7516	0.6834	0.6826	0.6660	0.6660
	1-2	0.8012	0.8012	0.7475	0.7462	0.6855	0.6855	0.6663	0.6662
	1-2-3	0.7959	0.7949	0.7500	0.7489	0.6861	0.6857	0.6681	0.6681
DIST_CHAR_ASIS_POS_TAG	1	0.7933	0.7920	0.7483	0.7469	0.6871	0.6871	0.6683	0.6682
	1-2	0.7972	0.7968	0.7484	0.7478	0.6863	0.6863	0.6694	0.6694
	1-2-3	0.7988	0.7988	0.7512	0.7511	0.6846	0.6828	0.6701	0.6698
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7954	0.7952	0.7530	0.7529	0.6861	0.6849	0.6700	0.6696
	1-2	0.7913	0.7912	0.7509	0.7507	0.6879	0.6879	0.6703	0.6701
	1-2-3	0.7939	0.7937	0.7534	0.7533	0.6877	0.6872	0.6713	0.6712
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7919	0.7918	0.7544	0.7544	0.6872	0.6858	0.6738	0.6738
	1-2	0.7929	0.7927	0.7559	0.7555	0.6876	0.6861	0.6739	0.6739
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7898	0.7895	0.7574	0.7573	0.6921	0.6921	0.6749	0.6747
	1-2	0.7883	0.7883	0.7510	0.7503	0.6913	0.6913	0.6751	0.6749

Table E.41: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1683	0.2888	-0.0348	-0.0259	-0.0312	-0.0785
	2-3	0.0446	0.0485	0.0274	0.0075	0.1002	0.0619
	2-3-4	0.0159	0.0471	0.0272	0.0942	0.0405	-0.0265
	2-3-4-5	-0.0486	-0.0061	-0.0090	-0.0183	0.0560	0.0661
DIST_CHAR	2	-0.0481	-0.0285	-0.0020	-0.0628	0.0686	0.0497
	2-3	-0.0355	0.0028	0.0120	0.0464	0.0665	0.0553
	2-3-4	-0.0022	0.0259	0.0037	0.0052	0.0198	0.0238
	2-3-4-5	-0.0103	0.0157	0.0228	0.0138	0.0165	0.0021
DIST_CHAR_ASIS	2	-0.0215	0.0133	-0.0312	0.0060	0.0065	0.0232
	2-3	-0.0055	-0.0018	0.0044	0.0263	-0.0050	0.0227
	2-3-4	-0.0265	-0.0274	0.0009	-0.0053	0.0245	0.0168
	2-3-4-5	-0.0284	-0.0212	-0.0118	-0.0198	-0.0143	-0.0140
DIST_CHAR_ASIS_POS	1	0.0138	0.0230	-0.0061	0.0064	0.0482	0.0538
	1-2	-0.0100	-0.0181	0.0198	0.0108	0.0833	0.0697
	1-2-3	-0.0009	0.0076	0.0354	0.0258	0.0346	0.0240
DIST_CHAR_ASIS_POS_TAG	1	-0.0063	0.0042	0.0069	-0.0115	0.0310	0.0311
	1-2	-0.0366	-0.0332	-0.0115	-0.0066	0.0082	0.0118
	1-2-3	-0.0141	-0.0133	0.0296	0.0455	0.0383	0.0351
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0154	-0.0017	0.0028	-0.0029	0.0159	0.0247
	1-2	-0.0077	0.0165	-0.0028	0.0099	0.0089	0.0170
	1-2-3	0.0008	0.0040	0.0022	0.0015	-0.0248	-0.0046
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0058	0.0071	0.0076	0.0101	0.0037	-0.0090
	1-2	-0.0004	0.0174	0.0129	-0.0012	0.0129	0.0178
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0031	0.0001	0.0124	0.0225	0.0171	0.0133
	1-2	0.0016	-0.0074	0.0278	0.0162	0.0380	0.0168

Table E.42: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

	Target	Age							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.5983	0.5994	0.4887	0.4861	0.3739	0.3555	0.3088	0.2673
	2-3	0.6907	0.6908	0.5595	0.5583	0.4424	0.4382	0.3888	0.3706
	2-3-4	0.7031	0.7037	0.5749	0.5714	0.4523	0.4435	0.4023	0.3817
	2-3-4-5	0.6804	0.6827	0.5378	0.5386	0.4624	0.4589	0.4017	0.3870
DIST_CHAR_ASIS	2	0.7161	0.7161	0.5828	0.5810	0.4677	0.4662	0.4223	0.4062
	2-3	0.7275	0.7276	0.5944	0.5926	0.4759	0.4726	0.4336	0.4241
	2-3-4	0.7270	0.7267	0.6014	0.6002	0.4744	0.4672	0.4309	0.4160
	2-3-4-5	0.7292	0.7295	0.5952	0.5939	0.4788	0.4738	0.4320	0.4203
DIST_CHAR_ASIS_POS	1	0.7229	0.7231	0.6008	0.5990	0.4822	0.4804	0.4324	0.4223
	1-2	0.7283	0.7283	0.5954	0.5942	0.4743	0.4679	0.4354	0.4266
	1-2-3	0.7053	0.7041	0.5854	0.5851	0.4850	0.4815	0.4392	0.4313
DIST_CHAR_ASIS_POS_TAG	1	0.7159	0.7162	0.5971	0.5952	0.4783	0.4756	0.4294	0.4164
	1-2	0.7004	0.6989	0.5946	0.5937	0.4823	0.4776	0.4364	0.4297
	1-2-3	0.6975	0.6957	0.5950	0.5931	0.4875	0.4843	0.4380	0.4277
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7124	0.7132	0.5923	0.5914	0.4856	0.4817	0.4367	0.4257
	1-2	0.7081	0.7076	0.5920	0.5910	0.4843	0.4809	0.4338	0.4204
	1-2-3	0.7075	0.7080	0.5952	0.5940	0.4865	0.4838	0.4382	0.4341
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.6861	0.6845	0.5950	0.5939	0.4885	0.4842	0.4362	0.4247
	1-2	0.7079	0.7082	0.5969	0.5956	0.4884	0.4844	0.4371	0.4269
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7058	0.7055	0.5969	0.5957	0.4890	0.4857	0.4349	0.4218
	1-2	0.7046	0.7043	0.5893	0.5879	0.4887	0.4856	0.4357	0.4228

Table E.43: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a cumulated model on the ordered, full feature set

	Target	Age					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0348	-0.0149	0.0099	0.0315	-0.0141	0.0174
	2-3	-0.0051	-0.0070	-0.0063	-0.0108	0.0126	0.0458
	2-3-4	-0.0033	0.0366	0.0245	-0.0002	0.0020	0.0207
	2-3-4-5	0.0213	0.0143	-0.0097	-0.0042	-0.0200	-0.0321
DIST_CHAR	2	0.0013	-0.0092	0.0077	0.0500	0.0047	0.0328
	2-3	0.0064	0.0009	-0.0084	-0.0053	0.0196	0.0099
	2-3-4	-0.0009	0.0049	0.0101	0.0123	-0.0004	-0.0037
	2-3-4-5	0.0157	0.0139	0.0107	-0.0099	0.0016	0.0008
DIST_CHAR_ASIS	2	0.0108	0.0058	-0.0019	0.0102	0.0090	0.0187
	2-3	0.0089	0.0029	0.0068	0.0036	0.0053	0.0122
	2-3-4	0.0119	0.0045	0.0045	0.0047	0.0087	0.0154
	2-3-4-5	0.0074	0.0087	-0.0049	-0.0032	0.0018	0.0034
DIST_CHAR_ASIS_POS	1	0.0141	0.0144	0.0166	0.0247	0.0187	0.0186
	1-2	0.0264	0.0325	0.0420	0.0360	0.0075	0.0030
	1-2-3	0.0127	0.0213	0.0150	0.0151	0.0125	0.0151
DIST_CHAR_ASIS_POS_TAG	1	0.0108	0.0118	0.0413	0.0368	0.0266	0.0279
	1-2	0.0067	0.0070	0.0243	0.0161	0.0117	0.0148
	1-2-3	0.0105	0.0177	0.0271	0.0232	0.0076	0.0183
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0139	0.0094	0.0129	0.0161	0.0093	0.0124
	1-2	0.0209	0.0192	0.0183	0.0169	0.0102	0.0107
	1-2-3	0.0229	0.0251	0.0072	0.0043	-0.0013	0.0013
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0245	0.0257	0.0189	0.0198	0.0154	0.0243
	1-2	0.0081	0.0135	0.0135	0.0140	0.0276	0.0304
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	-0.0017	-0.0044	0.0071	0.0104	0.0224	0.0262
	1-2	0.0203	0.0132	0.0189	0.0189	0.0241	0.0298

Minimum of Characters: 250 & Cumulated

Table E.44: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7844	0.7841	0.7495	0.7493	0.6736	0.6736	0.6514	0.6511
	2-3	0.8249	0.8247	0.7771	0.7738	0.7219	0.7218	0.6916	0.6916
	2-3-4	0.8310	0.8309	0.7884	0.7858	0.7306	0.7305	0.7001	0.7001
	2-3-4-5	0.8316	0.8315	0.7943	0.7927	0.7283	0.7279	0.6995	0.6995
DIST_CHAR_ASIS	2	0.8434	0.8434	0.8019	0.8002	0.7419	0.7418	0.7084	0.7084
	2-3	0.8423	0.8418	0.8120	0.8104	0.7475	0.7473	0.7166	0.7165
	2-3-4	0.8499	0.8499	0.8260	0.8260	0.7497	0.7495	0.7176	0.7172
	2-3-4-5	0.8479	0.8479	0.8243	0.8243	0.7485	0.7484	0.7173	0.7168
DIST_CHAR_ASIS_POS	1	0.8483	0.8483	0.8155	0.8145	0.7474	0.7470	0.7171	0.7171
	1-2	0.8439	0.8436	0.8170	0.8163	0.7486	0.7485	0.7184	0.7183
	1-2-3	0.8430	0.8430	0.8138	0.8128	0.7505	0.7505	0.7193	0.7193
DIST_CHAR_ASIS_POS_TAG	1	0.8441	0.8440	0.8159	0.8152	0.7472	0.7459	0.7176	0.7174
	1-2	0.8390	0.8389	0.8169	0.8168	0.7509	0.7509	0.7185	0.7184
	1-2-3	0.8377	0.8377	0.8134	0.8132	0.7521	0.7520	0.7185	0.7180
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8345	0.8342	0.8131	0.8126	0.7505	0.7505	0.7210	0.7210
	1-2	0.8345	0.8345	0.8152	0.8151	0.7513	0.7513	0.7169	0.7152
	1-2-3	0.8327	0.8327	0.8142	0.8141	0.7518	0.7516	0.7205	0.7203
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8298	0.8291	0.8173	0.8171	0.7429	0.7399	0.7178	0.7150
	1-2	0.8332	0.8328	0.8162	0.8161	0.7386	0.7340	0.7206	0.7186
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8294	0.8286	0.8133	0.8133	0.7559	0.7555	0.7232	0.7226
	1-2	0.8289	0.8286	0.8064	0.8051	0.7562	0.7558	0.7252	0.7251

Table E.45: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0390	0.1945	0.0539	0.0938	0.2805	0.2587
	2-3	0.1318	0.1969	0.0012	0.1452	0.1234	0.1113
	2-3-4	0.0671	0.1703	0.0232	0.0109	0.1353	0.0908
	2-3-4-5	0.0076	0.0844	0.0169	0.0175	0.0812	0.0675
DIST_CHAR	2	0.0017	0.0251	0.0348	0.0854	0.0363	0.0600
	2-3	-0.0013	0.0141	-0.0077	0.0357	0.0245	0.0117
	2-3-4	0.0015	0.0345	0.0257	0.0152	0.0389	0.0345
	2-3-4-5	-0.0060	0.0096	0.0168	0.0283	0.0403	0.0370
DIST_CHAR_ASIS	2	0.0181	0.0458	0.0196	-0.0173	-0.0216	-0.0147
	2-3	-0.0017	0.0099	0.0257	0.0034	0.0041	-0.0010
	2-3-4	0.0151	-0.0199	0.0072	-0.0106	0.0263	0.0332
	2-3-4-5	0.0060	0.0030	0.0127	-0.0121	0.0074	-0.0042
DIST_CHAR_ASIS_POS	1	0.0431	0.0190	-0.0015	-0.0213	0.0100	0.0060
	1-2	0.0327	0.0124	-0.0106	-0.0153	0.0417	0.0422
	1-2-3	0.0142	0.0324	0.0102	-0.0236	0.0124	0.0116
DIST_CHAR_ASIS_POS_TAG	1	-0.0230	-0.0049	0.0602	0.0363	0.0459	0.0400
	1-2	-0.0051	-0.0015	-0.0061	-0.0297	0.0271	0.0223
	1-2-3	0.0022	0.0010	-0.0143	-0.0176	0.0368	0.0398
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0084	0.0033	-0.0038	-0.0138	0.0160	0.0227
	1-2	0.0184	0.0234	0.0126	0.0124	0.0381	0.0358
	1-2-3	0.0184	0.0299	-0.0325	-0.0191	0.0387	0.0331
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0384	0.0402	-0.0095	-0.0214	0.0229	0.0133
	1-2	0.0124	0.0285	0.0307	0.0042	0.0130	0.0092
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0093	0.0051	0.0216	0.0340	0.0195	0.0214
	1-2	0.0252	0.0271	0.0238	0.0170	-0.0156	-0.0159

Table E.46: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

	Target	Age							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.6772	0.6770	0.5772	0.5772	0.4543	0.4439	0.3928	0.3686
	2-3	0.7627	0.7629	0.6684	0.6689	0.5239	0.5115	0.4643	0.4581
	2-3-4	0.7534	0.7555	0.6707	0.6732	0.5471	0.5440	0.4860	0.4795
	2-3-4-5	0.7380	0.7413	0.6847	0.6848	0.5454	0.5361	0.4927	0.4875
DIST_CHAR_ASIS	2	0.7749	0.7751	0.6993	0.6996	0.5607	0.5567	0.4827	0.4712
	2-3	0.7918	0.7920	0.7083	0.7067	0.5742	0.5705	0.5002	0.4942
	2-3-4	0.7981	0.7996	0.7140	0.7131	0.5773	0.5736	0.4982	0.4902
	2-3-4-5	0.7928	0.7941	0.7059	0.7071	0.5862	0.5845	0.5049	0.5020
DIST_CHAR_ASIS_POS	1	0.7622	0.7611	0.6947	0.6963	0.5817	0.5785	0.4990	0.4899
	1-2	0.7842	0.7846	0.7106	0.7104	0.5852	0.5838	0.4983	0.4921
	1-2-3	0.7857	0.7854	0.7052	0.7056	0.5856	0.5842	0.4992	0.4968
DIST_CHAR_ASIS_POS_TAG	1	0.7847	0.7851	0.7071	0.7068	0.5749	0.5713	0.4929	0.4799
	1-2	0.7752	0.7764	0.7012	0.7016	0.5699	0.5649	0.5002	0.4960
	1-2-3	0.7825	0.7829	0.7008	0.7012	0.5694	0.5635	0.5064	0.4999
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7744	0.7747	0.6959	0.6970	0.5797	0.5749	0.4959	0.4950
	1-2	0.7713	0.7721	0.6947	0.6960	0.5842	0.5829	0.4969	0.4925
	1-2-3	0.7823	0.7825	0.6981	0.6980	0.5832	0.5815	0.5061	0.5003
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7666	0.7668	0.6957	0.6956	0.5766	0.5709	0.5078	0.5031
	1-2	0.7642	0.7647	0.7027	0.7019	0.5735	0.5673	0.5108	0.5060
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7652	0.7674	0.6895	0.6883	0.5845	0.5821	0.5070	0.5042
	1-2	0.7686	0.7692	0.6882	0.6875	0.5860	0.5842	0.5078	0.5019

Table E.47: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a cumulated model on the ordered, full feature set

	Target	Age					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0416	0.1495	0.0072	0.0051	0.0153	0.0489
	2-3	-0.0223	0.0252	-0.0132	0.0114	0.0086	-0.0058
	2-3-4	0.0028	0.0019	0.0258	0.0180	0.0279	0.0153
	2-3-4-5	0.0040	0.0345	0.0270	0.0137	-0.0079	-0.0130
DIST_CHAR	2	0.0620	0.0267	0.0122	0.0015	0.0181	0.0090
	2-3	0.0686	0.0548	0.0189	0.0225	0.0152	0.0079
	2-3-4	0.0693	0.0148	0.0045	0.0058	0.0072	0.0068
	2-3-4-5	0.0496	0.0082	0.0093	0.0005	-0.0054	-0.0113
DIST_CHAR_ASIS	2	0.0390	0.0173	0.0187	0.0134	0.0050	0.0077
	2-3	0.0346	0.0165	0.0025	0.0088	0.0027	0.0066
	2-3-4	0.0333	0.0131	0.0004	0.0048	0.0168	0.0067
	2-3-4-5	0.0450	0.0078	-0.0004	0.0097	0.0206	-0.0040
DIST_CHAR_ASIS_POS	1	0.0487	0.0185	0.0226	0.0315	0.0268	0.0170
	1-2	0.0371	0.0085	0.0278	0.0285	0.0101	0.0091
	1-2-3	0.0359	0.0068	0.0253	0.0272	0.0106	0.0085
DIST_CHAR_ASIS_POS_TAG	1	0.0467	0.0368	0.0249	0.0209	0.0238	0.0231
	1-2	0.0567	0.0221	0.0192	0.0230	0.0132	0.0162
	1-2-3	0.0444	0.0179	0.0295	0.0261	0.0359	0.0319
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0169	-0.0045	0.0155	0.0171	0.0337	0.0358
	1-2	0.0212	0.0112	0.0235	0.0135	0.0254	0.0214
	1-2-3	0.0463	0.0107	0.0254	0.0261	0.0086	0.0097
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0495	0.0303	0.0171	0.0123	0.0171	0.0151
	1-2	0.0392	0.0134	0.0085	0.0048	0.0258	0.0210
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0342	0.0096	0.0089	0.0097	0.0285	0.0263
	1-2	0.0310	0.0073	0.0136	0.0129	0.0133	0.0160

Minimum of Characters: 500 & Cumulated

Table E.48: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.8231	0.8230	0.7868	0.7868	0.7062	0.7058	0.6795	0.6769
	2-3	0.8802	0.8802	0.8396	0.8392	0.7698	0.7696	0.7313	0.7311
	2-3-4	0.8727	0.8712	0.8498	0.8494	0.7852	0.7852	0.7407	0.7392
	2-3-4-5	0.8953	0.8951	0.8583	0.8583	0.7679	0.7650	0.7433	0.7429
DIST_CHAR_ASIS	2	0.8887	0.8882	0.8623	0.8623	0.7521	0.7441	0.7511	0.7494
	2-3	0.8861	0.8850	0.8725	0.8725	0.8036	0.8036	0.7612	0.7602
	2-3-4	0.8779	0.8764	0.8734	0.8733	0.7700	0.7641	0.7676	0.7675
	2-3-4-5	0.8947	0.8942	0.8730	0.8729	0.7688	0.7626	0.7674	0.7673
DIST_CHAR_ASIS_POS	1	0.9048	0.9045	0.8717	0.8717	0.8043	0.8042	0.7660	0.7656
	1-2	0.8976	0.8975	0.8670	0.8668	0.8040	0.8040	0.7678	0.7677
	1-2-3	0.8937	0.8936	0.8701	0.8700	0.7985	0.7976	0.7660	0.7650
DIST_CHAR_ASIS_POS_TAG	1	0.8996	0.8996	0.8696	0.8696	0.7980	0.7971	0.7687	0.7687
	1-2	0.8881	0.8881	0.8678	0.8677	0.7755	0.7704	0.7695	0.7694
	1-2-3	0.8792	0.8786	0.8640	0.8639	0.8058	0.8057	0.7692	0.7691
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8740	0.8729	0.8676	0.8676	0.8056	0.8055	0.7698	0.7698
	1-2	0.8615	0.8596	0.8634	0.8633	0.8037	0.8031	0.7698	0.7697
	1-2-3	0.8677	0.8664	0.8615	0.8615	0.7804	0.7761	0.7710	0.7709
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8835	0.8835	0.8628	0.8628	0.8050	0.8049	0.7754	0.7754
	1-2	0.8786	0.8786	0.8594	0.8591	0.7841	0.7801	0.7707	0.7695
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8704	0.8698	0.8639	0.8636	0.8057	0.8055	0.7759	0.7757
	1-2	0.8819	0.8814	0.8644	0.8644	0.8062	0.8060	0.7769	0.7767

Table E.49: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.0858	0.1370	0.1528	0.2654	0.2834	0.2834
	2-3	0.1535	0.1134	0.1350	0.1049	0.1451	0.1875
	2-3-4	0.1185	-0.0050	0.0893	0.0610	0.0425	0.0354
	2-3-4-5	0.0904	0.0550	0.0491	0.0644	0.0457	0.0703
DIST_CHAR	2	0.0911	0.0265	0.0598	0.0466	0.0998	0.0889
	2-3	0.0903	-0.0365	0.0331	0.0278	0.0178	0.0419
	2-3-4	0.0990	0.0314	0.0445	0.0261	0.0454	0.0362
	2-3-4-5	0.0829	-0.0121	0.0455	0.0365	0.0461	0.0232
DIST_CHAR_ASIS	2	0.0815	-0.0217	0.0306	0.0077	0.0118	0.0141
	2-3	0.0931	0.0012	0.0462	0.0171	0.0295	0.0138
	2-3-4	0.0789	-0.0137	0.0380	0.0166	0.0026	0.0088
	2-3-4-5	0.0829	0.0014	0.0459	0.0050	0.0281	0.0243
DIST_CHAR_ASIS_POS	1	0.1044	0.0228	0.0343	-0.0083	0.0389	0.0358
	1-2	0.0661	0.0052	-0.0052	-0.0298	0.0173	0.0290
	1-2-3	0.0603	0.0113	0.0389	0.0158	0.0362	0.0372
DIST_CHAR_ASIS_POS_TAG	1	0.0473	-0.0021	0.0340	0.0171	0.0295	0.0218
	1-2	0.0578	-0.0141	0.0384	0.0072	-0.0040	-0.0065
	1-2-3	0.0267	-0.0247	0.0100	-0.0060	0.0340	0.0434
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0737	0.0128	0.0054	-0.0020	0.0309	0.0268
	1-2	0.0797	0.0354	-0.0123	-0.0162	0.0159	0.0262
	1-2-3	0.0488	0.0061	0.0163	-0.0007	0.0464	0.0227
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0535	-0.0062	0.0102	-0.0180	0.0378	0.0332
	1-2	0.0436	0.0064	0.0343	0.0316	0.0089	0.0107
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0380	-0.0058	0.0038	-0.0200	0.0358	0.0473
	1-2	0.0617	0.0180	0.0094	0.0061	0.0258	0.0262

Table E.50: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

	Target	Age							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7262	0.7271	0.6350	0.6349	0.5133	0.5123	0.4465	0.4434
	2-3	0.8192	0.8214	0.7342	0.7331	0.5938	0.5898	0.5199	0.5177
	2-3-4	0.8321	0.8343	0.7623	0.7627	0.6151	0.6114	0.5437	0.5390
	2-3-4-5	0.8232	0.8250	0.7590	0.7595	0.6186	0.6140	0.5451	0.5397
DIST_CHAR_ASIS	2	0.8508	0.8518	0.7725	0.7727	0.6203	0.6150	0.5566	0.5515
	2-3	0.8727	0.8726	0.7954	0.7954	0.6640	0.6629	0.5724	0.5687
	2-3-4	0.8753	0.8753	0.8017	0.8014	0.6618	0.6610	0.5643	0.5553
	2-3-4-5	0.8673	0.8671	0.8007	0.8006	0.6421	0.6376	0.5764	0.5731
DIST_CHAR_ASIS_POS	1	0.8633	0.8632	0.7846	0.7846	0.6552	0.6541	0.5766	0.5736
	1-2	0.8736	0.8733	0.7838	0.7841	0.6565	0.6545	0.5797	0.5772
	1-2-3	0.8451	0.8466	0.7900	0.7893	0.6411	0.6362	0.5611	0.5498
DIST_CHAR_ASIS_POS_TAG	1	0.8526	0.8531	0.7741	0.7742	0.6620	0.6614	0.5714	0.5656
	1-2	0.8575	0.8579	0.7851	0.7850	0.6645	0.6640	0.5764	0.5727
	1-2-3	0.8362	0.8383	0.7776	0.7770	0.6530	0.6500	0.5609	0.5498
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8388	0.8391	0.7795	0.7792	0.6606	0.6606	0.5829	0.5806
	1-2	0.8482	0.8487	0.7723	0.7717	0.6534	0.6517	0.5753	0.5719
	1-2-3	0.8388	0.8399	0.7694	0.7683	0.6360	0.6308	0.5651	0.5569
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8495	0.8486	0.7737	0.7735	0.6416	0.6376	0.5705	0.5625
	1-2	0.8531	0.8526	0.7795	0.7790	0.6531	0.6509	0.5803	0.5756
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8353	0.8368	0.7718	0.7715	0.6654	0.6651	0.5751	0.5708
	1-2	0.8575	0.8575	0.7755	0.7751	0.6598	0.6591	0.5676	0.5600

Table E.51: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a cumulated model on the ordered, full feature set

	Target	Age					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST	2	0.1872	0.0727	0.0083	0.0109	0.0192	0.0501
	2-3	0.1142	0.0359	0.0275	0.0294	0.0478	0.0012
	2-3-4	0.0268	0.0157	0.0502	0.0416	0.0099	0.0147
	2-3-4-5	0.0219	0.0064	0.0371	0.0096	0.0154	0.0220
DIST_CHAR	2	0.1366	0.0099	0.0244	0.0326	0.0048	0.0061
	2-3	0.1364	0.0224	0.0286	0.0257	0.0117	0.0144
	2-3-4	0.1078	0.0159	0.0183	0.0196	0.0110	-0.0038
	2-3-4-5	0.0934	0.0227	0.0179	0.0167	0.0084	0.0023
DIST_CHAR_ASIS	2	0.0901	-0.0056	0.0189	0.0107	0.0071	0.0017
	2-3	0.0815	0.0144	0.0157	0.0055	0.0102	0.0100
	2-3-4	0.0848	0.0158	0.0124	0.0120	0.0053	0.0009
	2-3-4-5	0.0831	0.0126	0.0198	0.0020	0.0067	0.0095
DIST_CHAR_ASIS_POS	1	0.0836	0.0140	0.0336	0.0287	0.0294	0.0322
	1-2	0.0750	0.0146	0.0199	0.0174	0.0331	0.0223
	1-2-3	0.0830	0.0122	0.0178	0.0230	0.0221	0.0137
DIST_CHAR_ASIS_POS_TAG	1	0.0816	0.0207	0.0313	0.0318	0.0364	0.0321
	1-2	0.0806	0.0242	0.0142	0.0162	0.0363	0.0337
	1-2-3	0.0773	0.0271	0.0191	0.0185	0.0177	0.0214
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0728	0.0171	0.0158	0.0135	0.0362	0.0267
	1-2	0.0760	0.0222	0.0086	0.0084	0.0380	0.0439
	1-2-3	0.0780	0.0198	0.0306	0.0297	0.0268	0.0312
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0704	0.0164	0.0138	0.0153	0.0159	0.0129
	1-2	0.0574	0.0144	0.0129	0.0065	0.0227	0.0188
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0544	0.0200	0.0196	0.0187	0.0297	0.0340
	1-2	0.0652	0.0201	0.0148	0.0152	0.0180	0.0197

Minimum of Characters: 100 & Stacked

Table E.52: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	100							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7223	0.7220	0.6879	0.6864	0.6232	0.6230	0.6142	0.6136
	2-3	0.7658	0.7658	0.7159	0.7139	0.6594	0.6594	0.6426	0.6424
	2-3-4	0.7776	0.7776	0.7259	0.7253	0.6655	0.6655	0.6478	0.6475
	2-3-4-5	0.7732	0.7731	0.7249	0.7243	0.6638	0.6637	0.6477	0.6474
DIST_CHAR_ASIS	2	0.7761	0.7760	0.7217	0.7200	0.6657	0.6656	0.6487	0.6484
	2-3	0.7881	0.7880	0.7314	0.7300	0.6694	0.6693	0.6518	0.6516
	2-3-4	0.7914	0.7914	0.7361	0.7348	0.6723	0.6723	0.6543	0.6540
	2-3-4-5	0.7890	0.7889	0.7338	0.7324	0.6708	0.6707	0.6540	0.6539
DIST_CHAR_ASIS_POS	1	0.7893	0.7893	0.7334	0.7318	0.6709	0.6708	0.6535	0.6532
	1-2	0.7889	0.7888	0.7340	0.7325	0.6708	0.6707	0.6533	0.6530
	1-2-3	0.7903	0.7903	0.7335	0.7320	0.6711	0.6711	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG	1	0.7889	0.7888	0.7340	0.7324	0.6721	0.6721	0.6546	0.6542
	1-2	0.7898	0.7897	0.7329	0.7314	0.6709	0.6709	0.6537	0.6534
	1-2-3	0.7900	0.7900	0.7329	0.7313	0.6717	0.6716	0.6538	0.6535
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7890	0.7889	0.7335	0.7319	0.6724	0.6724	0.6541	0.6539
	1-2	0.7892	0.7891	0.7295	0.7273	0.6720	0.6720	0.6537	0.6534
	1-2-3	0.7911	0.7910	0.7324	0.7306	0.6720	0.6720	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7907	0.7906	0.7326	0.7310	0.6718	0.6718	0.6539	0.6536
	1-2	0.7912	0.7911	0.7328	0.7311	0.6718	0.6718	0.6537	0.6534
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7927	0.7926	0.7340	0.7324	0.6733	0.6733	0.6552	0.6549
	1-2	0.7924	0.7923	0.7336	0.7320	0.6724	0.6723	0.6550	0.6548

Table E.53: Stability of feature relevance for the prediction of gender on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	100					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	-0.0387	-0.0005	0.0055	-0.0134	0.0438	0.0420
	2-3	-0.0308	-0.0107	-0.0109	-0.0139	0.0442	0.0409
	2-3-4	-0.0293	-0.0093	-0.0037	-0.0112	0.0369	0.0404
	2-3-4-5	-0.0168	0.0059	-0.0041	-0.0094	0.0306	0.0362
DIST_CHAR_ASIS	2	-0.0102	0.0104	-0.0034	-0.0049	0.0243	0.0322
	2-3	-0.0078	0.0036	0.0263	-0.0042	-0.0069	0.0140
	2-3-4	-0.0236	-0.0164	-0.0019	-0.0071	0.0207	0.0259
	2-3-4-5	-0.0155	-0.0105	-0.0049	-0.0039	0.0203	0.0195
DIST_CHAR_ASIS_POS	1	-0.0262	-0.0216	-0.0101	-0.0092	0.0001	-0.0007
	1-2	-0.0185	-0.0130	0.0032	0.0068	0.0391	0.0421
	1-2-3	-0.0077	0.0013	-0.0343	-0.0375	0.0188	0.0232
DIST_CHAR_ASIS_POS_TAG	1	-0.0131	-0.0047	-0.0303	-0.0333	0.0116	0.0157
	1-2	-0.0179	-0.0111	-0.0474	-0.0499	0.0216	0.0252
	1-2-3	-0.0209	-0.0095	-0.0277	-0.0273	0.0166	0.0257
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0184	-0.0077	-0.0274	-0.0269	0.0178	0.0263
	1-2	-0.0330	-0.0194	-0.0215	-0.0223	0.0134	0.0237
	1-2-3	-0.0284	-0.0184	-0.0342	-0.0277	0.0249	0.0359
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	-0.0257	-0.0151	-0.0325	-0.0235	0.0232	0.0351
	1-2	0.0076	0.0090	-0.0325	-0.0245	0.0215	0.0322
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0084	0.0118	-0.0307	-0.0221	0.0209	0.0316
	1-2	0.0082	0.0058	-0.0320	-0.0221	0.0256	0.0320

Table E.54: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 100 No. of Authors: 50 Score N-gram ranges	Accuracy	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.5814	0.5806	0.4549	0.4544	0.2092	0.0892	0.2029	0.0707
	2-3	0.6560	0.6565	0.5434	0.5426	0.4313	0.4281	0.2607	0.1626
	2-3-4	0.6726	0.6728	0.5630	0.5621	0.4346	0.4147	0.3547	0.2988
	2-3-4-5	0.6769	0.6773	0.5647	0.5638	0.4531	0.4464	0.3450	0.2828
DIST_CHAR_ASIS	2	0.6955	0.6954	0.5636	0.5627	0.4507	0.4473	0.3738	0.3267
	2-3	0.7126	0.7128	0.5758	0.5750	0.4604	0.4584	0.3820	0.3378
	2-3-4	0.7164	0.7163	0.5829	0.5822	0.4703	0.4649	0.3707	0.3001
	2-3-4-5	0.7156	0.7156	0.5835	0.5828	0.4695	0.4638	0.3786	0.3584
DIST_CHAR_ASIS_POS	1	0.7158	0.7159	0.5835	0.5828	0.4693	0.4635	0.3673	0.3417
	1-2	0.7158	0.7159	0.5826	0.5819	0.4693	0.4634	0.3673	0.3417
	1-2-3	0.7163	0.7164	0.5831	0.5825	0.4693	0.4635	0.3679	0.3423
DIST_CHAR_ASIS_POS_TAG	1	0.7162	0.7163	0.5832	0.5825	0.4696	0.4635	0.3831	0.3488
	1-2	0.7159	0.7161	0.5832	0.5825	0.4695	0.4634	0.3833	0.3491
	1-2-3	0.7183	0.7184	0.5835	0.5828	0.4695	0.4635	0.3831	0.3489
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7183	0.7184	0.5832	0.5825	0.4569	0.4440	0.3812	0.3464
	1-2	0.7185	0.7186	0.5833	0.5826	0.4697	0.4634	0.3812	0.3465
	1-2-3	0.7187	0.7189	0.5833	0.5826	0.4697	0.4634	0.3812	0.3464
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7179	0.7184	0.5842	0.5835	0.4703	0.4638	0.3793	0.3440
	1-2	0.7191	0.7192	0.5826	0.5820	0.4704	0.4639	0.3797	0.3443
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7197	0.7198	0.5851	0.5843	0.4706	0.4639	0.4036	0.3761
	1-2	0.7196	0.7197	0.5853	0.5846	0.4705	0.4639	0.4019	0.3724

Table E.55: Stability of feature relevance for the prediction of age on a minimal input instance length of 100 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 100 No. of Authors: 150 Score N-gram ranges	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	-0.0261	-0.0180	-0.0041	0.0083	0.0156	0.0209
	2-3	-0.0058	-0.0031	-0.0170	-0.0032	0.0285	0.0297
	2-3-4	-0.0065	0.0028	0.0036	0.0192	-0.0086	-0.0014
	2-3-4-5	-0.0145	-0.0003	-0.0099	0.0049	0.0225	0.0185
DIST_CHAR_ASIS	2	0.0049	-0.0039	0.0051	0.0080	0.0112	0.0043
	2-3	-0.0002	-0.0051	-0.0048	0.0081	0.0121	0.0122
	2-3-4	0.0066	0.0059	-0.0038	0.0006	0.0089	0.0184
	2-3-4-5	0.0106	0.0062	-0.0028	0.0105	0.0121	0.0134
DIST_CHAR_ASIS_POS	1	0.0128	0.0108	0.0211	0.0260	0.0288	0.0307
	1-2	0.0068	0.0038	0.0031	0.0109	0.0196	0.0211
	1-2-3	0.0035	-0.0061	0.0233	0.0247	0.0061	0.0055
DIST_CHAR_ASIS_POS_TAG	1	0.0132	0.0031	0.0259	0.0330	0.0098	0.0031
	1-2	0.0141	0.0051	0.0348	0.0395	0.0161	0.0098
	1-2-3	0.0125	-0.0036	0.0332	0.0297	0.0227	0.0170
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0149	0.0027	0.0260	0.0241	0.0154	0.0133
	1-2	0.0281	0.0163	0.0388	0.0381	0.0344	0.0309
	1-2-3	0.0178	0.0115	0.0328	0.0333	0.0331	0.0304
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0309	0.0213	0.0342	0.0276	0.0324	0.0308
	1-2	0.0239	0.0197	0.0286	0.0290	0.0267	0.0261
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0181	0.0120	0.0313	0.0288	0.0259	0.0256
	1-2	0.0074	0.0104	0.0267	0.0247	0.0264	0.0268

Minimum of Characters: 250 & Stacked

Table E.56: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	250							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.7737	0.7735	0.7414	0.7414	0.6647	0.6642	0.6455	0.6436
	2-3	0.8138	0.8130	0.7813	0.7813	0.7156	0.7156	0.6858	0.6849
	2-3-4	0.8218	0.8217	0.7912	0.7911	0.7261	0.7261	0.6933	0.6926
	2-3-4-5	0.8230	0.8230	0.7927	0.7926	0.7266	0.7266	0.6936	0.6928
DIST_CHAR_ASIS	2	0.8287	0.8287	0.7983	0.7983	0.7282	0.7280	0.6947	0.6940
	2-3	0.8397	0.8396	0.8076	0.8076	0.7323	0.7323	0.6996	0.6990
	2-3-4	0.8392	0.8391	0.8108	0.8106	0.7353	0.7353	0.7051	0.7050
	2-3-4-5	0.8359	0.8357	0.8114	0.8113	0.7345	0.7345	0.7046	0.7045
DIST_CHAR_ASIS_POS	1	0.8218	0.8200	0.8128	0.8127	0.7345	0.7345	0.7041	0.7040
	1-2	0.8401	0.8400	0.8095	0.8092	0.7348	0.7348	0.7042	0.7040
	1-2-3	0.8419	0.8419	0.8105	0.8104	0.7342	0.7342	0.7034	0.7027
DIST_CHAR_ASIS_POS_TAG	1	0.8401	0.8399	0.8127	0.8126	0.7357	0.7357	0.7040	0.7034
	1-2	0.8437	0.8437	0.8094	0.8090	0.7347	0.7347	0.7038	0.7037
	1-2-3	0.8105	0.8085	0.8104	0.8103	0.7348	0.7348	0.7034	0.7028
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8417	0.8417	0.8085	0.8082	0.7362	0.7362	0.7031	0.7026
	1-2	0.8376	0.8375	0.8082	0.8079	0.7350	0.7350	0.7038	0.7037
	1-2-3	0.8376	0.8374	0.8105	0.8104	0.7351	0.7351	0.7026	0.7021
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8387	0.8386	0.8072	0.8069	0.7330	0.7330	0.7027	0.7022
	1-2	0.8397	0.8396	0.8105	0.8104	0.7351	0.7351	0.7027	0.7021
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8365	0.8363	0.8067	0.8064	0.7363	0.7359	0.7042	0.7036
	1-2	0.8383	0.8381	0.8054	0.8050	0.7364	0.7361	0.7054	0.7054

Table E.57: Stability of feature relevance for the prediction of gender on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	250					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.0315	0.0538	0.0180	0.0214	0.0739	0.0638
	2-3	0.0250	0.0498	0.0184	0.0121	0.0694	0.0464
	2-3-4	0.0295	0.0412	0.0155	0.0154	0.0480	0.0440
	2-3-4-5	0.0248	0.0451	0.0103	0.0112	0.0500	0.0452
DIST_CHAR_ASIS	2	0.0336	0.0458	0.0110	0.0130	0.0486	0.0443
	2-3	0.0315	0.0388	0.0124	0.0391	0.0262	0.0248
	2-3-4	0.0327	0.0429	0.0126	0.0066	0.0414	0.0400
	2-3-4-5	0.0557	-0.0074	0.0112	0.0075	0.0403	0.0336
DIST_CHAR_ASIS_POS	1	0.0237	0.0340	0.0405	0.0371	0.0685	0.0623
	1-2	0.0525	0.0020	0.0072	0.0242	0.0643	0.0586
	1-2-3	0.0106	0.0140	-0.0034	-0.0059	0.0298	0.0240
DIST_CHAR_ASIS_POS_TAG	1	0.0069	0.0149	0.0095	0.0072	0.0283	0.0229
	1-2	0.0019	-0.0313	0.0135	0.0283	0.0243	0.0202
	1-2-3	0.0152	0.0117	-0.0018	-0.0021	0.0263	0.0163
DIST_CHAR_ASIS_POS_TAG_DEP	1	-0.0049	-0.0251	0.0210	0.0162	0.0381	0.0286
	1-2	-0.0067	-0.0289	0.0067	0.0067	0.0194	0.0124
	1-2-3	0.0186	0.0083	-0.0189	-0.0083	0.0308	0.0178
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0143	0.0056	-0.0178	-0.0085	0.0274	0.0191
	1-2	0.0173	0.0120	-0.0170	-0.0071	0.0277	0.0177
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0133	0.0108	-0.0146	-0.0058	0.0275	0.0180
	1-2	0.0136	0.0149	-0.0143	-0.0054	0.0260	0.0156

Table E.58: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 250 No. of Authors / Score / N-gram ranges	50 Accuracy	50 F1-score	150 Accuracy	150 F1-score	500 Accuracy	500 F1-score	1000 Accuracy	1000 F1-score
DIST_CHAR	2	0.6667	0.6672	0.5623	0.5621	0.4216	0.4205	0.3466	0.3196
	2-3	0.7480	0.7484	0.6506	0.6498	0.5183	0.5180	0.4535	0.4513
	2-3-4	0.7588	0.7592	0.6741	0.6735	0.5337	0.5308	0.4774	0.4757
	2-3-4-5	0.7627	0.7630	0.6722	0.6718	0.5325	0.5292	0.4829	0.4793
DIST_CHAR_ASIS	2	0.7759	0.7763	0.6793	0.6788	0.5455	0.5419	0.4849	0.4810
	2-3	0.7940	0.7943	0.6919	0.6914	0.5571	0.5567	0.4930	0.4892
	2-3-4	0.7911	0.7915	0.6996	0.6991	0.5553	0.5529	0.4994	0.4962
	2-3-4-5	0.7933	0.7933	0.6966	0.6962	0.5553	0.5530	0.5000	0.4972
DIST_CHAR_ASIS_POS	1	0.7869	0.7874	0.6967	0.6962	0.5553	0.5530	0.5000	0.4975
	1-2	0.7930	0.7930	0.6969	0.6965	0.5553	0.5529	0.5001	0.4976
	1-2-3	0.7986	0.7990	0.6975	0.6971	0.5554	0.5530	0.5000	0.4975
DIST_CHAR_ASIS_POS_TAG	1	0.7984	0.7989	0.6974	0.6970	0.5547	0.5522	0.4973	0.4965
	1-2	0.7986	0.7988	0.6965	0.6963	0.5547	0.5523	0.4973	0.4965
	1-2-3	0.7984	0.7990	0.6981	0.6974	0.5550	0.5525	0.4975	0.4966
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.7979	0.7985	0.6976	0.6969	0.5625	0.5625	0.4974	0.4967
	1-2	0.7986	0.7992	0.6978	0.6970	0.5626	0.5626	0.4975	0.4968
	1-2-3	0.7981	0.7985	0.6986	0.6979	0.5628	0.5628	0.4976	0.4969
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.7893	0.7899	0.6971	0.6964	0.5560	0.5537	0.4974	0.4968
	1-2	0.7925	0.7930	0.6966	0.6959	0.5554	0.5531	0.4972	0.4967
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.7962	0.7964	0.6967	0.6960	0.5604	0.5568	0.4972	0.4967
	1-2	0.8003	0.8005	0.6973	0.6966	0.5565	0.5544	0.4974	0.4969

Table E.59: Stability of feature relevance for the prediction of age on a minimal input instance length of 250 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 250 No. of Authors: 150 Score / N-gram ranges	500 Avg. Spearman's ρ (ext.)	500 Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	1000 Avg. Spearman's ρ (red.)	1500 Avg. Spearman's ρ (ext.)	1500 Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0132	-0.0179	-0.0008	0.0021	0.0057	0.0131
	2-3	0.0076	-0.0155	0.0017	0.0131	0.0161	0.0122
	2-3-4	0.0248	0.0128	-0.0031	0.0107	0.0037	0.0109
	2-3-4-5	0.0154	—	0.0055	0.0092	0.0169	0.0218
DIST_CHAR_ASIS	2	0.0218	0.0024	-0.0060	-0.0045	0.0118	0.0127
	2-3	0.0311	-0.0075	0.0108	0.0040	0.0065	0.0082
	2-3-4	0.0289	-0.0034	0.0092	0.0006	-0.0009	-0.0029
	2-3-4-5	0.0335	0.0136	0.0001	0.0041	0.0002	—
DIST_CHAR_ASIS_POS	1	0.0386	0.0154	0.0174	0.0131	0.0050	0.0073
	1-2	0.0268	0.0082	0.0128	0.0042	0.0088	0.0037
	1-2-3	0.0241	0.0055	-0.0044	-0.0052	0.0085	0.0093
DIST_CHAR_ASIS_POS_TAG	1	0.0239	0.0045	-0.0050	-0.0100	0.0033	0.0047
	1-2	0.0252	0.0051	0.0043	-0.0025	0.0144	0.0131
	1-2-3	0.0344	0.0300	-0.0073	-0.0073	0.0085	0.0086
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0329	0.0234	-0.0059	-0.0065	0.0145	0.0159
	1-2	0.0298	0.0162	-0.0067	-0.0084	0.0237	0.0231
	1-2-3	0.0259	0.0234	-0.0017	0.0031	0.0254	0.0271
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0212	0.0223	0.0092	0.0033	0.0190	0.0194
	1-2	0.0212	0.0179	0.0037	0.0047	0.0215	0.0199
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0212	0.0178	0.0053	-0.0020	0.0219	0.0186
	1-2	0.0144	0.0180	0.0050	0.0030	0.0124	0.0166

Minimum of Characters: 500 & Stacked

Table E.60: Accuracy and F1-scores for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

	Target	Gender							
	Min. No. of Characters	500							
	No. of Authors	50		150		500		1000	
	Score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Feature types	N-gram ranges								
DIST_CHAR	2	0.8192	0.8179	0.7780	0.7778	0.6948	0.6935	0.6765	0.6764
	2-3	0.8743	0.8740	0.8369	0.8369	0.7639	0.7636	0.7286	0.7285
	2-3-4	0.8874	0.8874	0.8485	0.8485	0.7772	0.7769	0.7400	0.7399
	2-3-4-5	0.8884	0.8883	0.8501	0.8500	0.7773	0.7769	0.7411	0.7410
DIST_CHAR_ASIS	2	0.8920	0.8919	0.8527	0.8527	0.7767	0.7761	0.7412	0.7411
	2-3	0.8983	0.8982	0.8619	0.8619	0.7844	0.7841	0.7467	0.7466
	2-3-4	0.8832	0.8830	0.8670	0.8670	0.7862	0.7855	0.7505	0.7504
	2-3-4-5	0.8789	0.8787	0.8664	0.8663	0.7856	0.7850	0.7515	0.7513
DIST_CHAR_ASIS_POS	1	0.8756	0.8754	0.8653	0.8652	0.7858	0.7852	0.7526	0.7525
	1-2	0.8986	0.8982	0.8673	0.8672	0.7860	0.7854	0.7516	0.7515
	1-2-3	0.8819	0.8817	0.8663	0.8662	0.7860	0.7854	0.7513	0.7512
DIST_CHAR_ASIS_POS_TAG	1	0.8815	0.8814	0.8666	0.8665	0.7851	0.7845	0.7522	0.7522
	1-2	0.8940	0.8940	0.8659	0.8659	0.7861	0.7855	0.7513	0.7511
	1-2-3	0.8924	0.8923	0.8664	0.8663	0.7864	0.7859	0.7514	0.7512
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8897	0.8897	0.8647	0.8646	0.7878	0.7873	0.7525	0.7525
	1-2	0.8871	0.8871	0.8664	0.8663	0.7875	0.7869	0.7515	0.7514
	1-2-3	0.8933	0.8933	0.8665	0.8664	0.7864	0.7859	0.7511	0.7509
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8858	0.8854	0.8674	0.8673	0.7865	0.7859	0.7511	0.7510
	1-2	0.8989	0.8987	0.8672	0.8671	0.7862	0.7856	0.7509	0.7508
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8828	0.8822	0.8673	0.8672	0.7872	0.7866	0.7523	0.7523
	1-2	0.8966	0.8963	0.8684	0.8684	0.7872	0.7866	0.7514	0.7514

Table E.61: Stability of feature relevance for the prediction of gender on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

	Target	Gender					
	Min. No. of Characters	500					
	No. of Authors	150		500		1000	
	Score	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
Feature types	N-gram ranges						
DIST_CHAR	2	0.1274	0.0535	0.0573	0.0718	0.0552	0.0749
	2-3	0.1182	0.0385	0.0420	0.0543	0.0388	0.0574
	2-3-4	0.1085	0.0315	0.0355	0.0490	0.0417	0.0531
	2-3-4-5	0.0951	0.0258	0.0311	0.0432	0.0336	0.0474
DIST_CHAR_ASIS	2	0.1148	0.0277	0.0340	0.0439	0.0364	0.0494
	2-3	0.1371	-0.0178	0.0286	0.0425	0.0308	0.0454
	2-3-4	0.1046	0.0192	0.0263	0.0358	0.0282	0.0381
	2-3-4-5	0.0987	0.0262	0.0281	0.0320	0.0272	0.0380
DIST_CHAR_ASIS_POS	1	0.1100	0.0431	0.0500	0.0536	0.0403	0.0503
	1-2	0.0886	0.0270	0.0025	0.0091	0.0610	0.0674
	1-2-3	0.0702	0.0068	0.0181	0.0232	0.0334	0.0421
DIST_CHAR_ASIS_POS_TAG	1	0.0644	0.0050	0.0067	0.0122	0.0316	0.0398
	1-2	0.0779	0.0190	0.0207	0.0189	0.0429	0.0500
	1-2-3	0.0695	0.0170	0.0159	0.0142	0.0485	0.0553
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0664	0.0061	0.0203	0.0188	0.0472	0.0537
	1-2	0.0733	0.0239	0.0040	0.0025	0.0411	0.0456
	1-2-3	0.0548	0.0065	0.0150	0.0154	0.0521	0.0573
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0522	-0.0209	0.0153	0.0168	0.0516	0.0543
	1-2	0.0481	0.0099	0.0146	0.0128	0.0516	0.0498
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0449	-0.0142	0.0145	0.0117	0.0515	0.0499
	1-2	0.0431	-0.0145	0.0179	0.0132	0.0499	0.0451

Table E.62: Accuracy and F1-scores for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 500 No. of Authors: 50 Score: Accuracy N-gram ranges	F1-score	150 Accuracy	F1-score	500 Accuracy	F1-score	1000 Accuracy	F1-score
DIST_CHAR	2	0.7413	0.7422	0.6347	0.6350	0.4899	0.4848	0.4263	0.4239
	2-3	0.8250	0.8252	0.7357	0.7356	0.5868	0.5840	0.5140	0.5114
	2-3-4	0.8415	0.8416	0.7623	0.7622	0.6156	0.6141	0.5433	0.5432
	2-3-4-5	0.8397	0.8397	0.7605	0.7603	0.6172	0.6161	0.5433	0.5416
DIST_CHAR_ASIS	2	0.8473	0.8472	0.7703	0.7701	0.6201	0.6187	0.5489	0.5469
	2-3	0.8651	0.8652	0.7795	0.7795	0.6280	0.6258	0.5550	0.5537
	2-3-4	0.8749	0.8748	0.7912	0.7911	0.6400	0.6396	0.5630	0.5630
	2-3-4-5	0.8700	0.8700	0.7878	0.7876	0.6372	0.6357	0.5623	0.5623
DIST_CHAR_ASIS_POS	1	0.8713	0.8709	0.7878	0.7877	0.6375	0.6360	0.5621	0.5621
	1-2	0.8718	0.8715	0.7881	0.7879	0.6375	0.6360	0.5621	0.5621
	1-2-3	0.8749	0.8746	0.7870	0.7868	0.6372	0.6357	0.5620	0.5621
DIST_CHAR_ASIS_POS_TAG	1	0.8767	0.8767	0.7869	0.7867	0.6395	0.6385	0.5621	0.5621
	1-2	0.8793	0.8793	0.7843	0.7833	0.6395	0.6385	0.5636	0.5636
	1-2-3	0.8820	0.8821	0.7910	0.7908	0.6417	0.6415	0.5628	0.5629
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.8802	0.8803	0.7905	0.7904	0.6419	0.6418	0.5626	0.5627
	1-2	0.8736	0.8735	0.7911	0.7909	0.6376	0.6351	0.5626	0.5627
	1-2-3	0.8793	0.8794	0.7912	0.7911	0.6418	0.6417	0.5631	0.5632
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.8740	0.8741	0.7911	0.7909	0.6387	0.6362	0.5632	0.5633
	1-2	0.8776	0.8773	0.7914	0.7912	0.6387	0.6361	0.5633	0.5634
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.8758	0.8756	0.7946	0.7945	0.6399	0.6385	0.5634	0.5635
	1-2	0.8762	0.8761	0.7911	0.7908	0.6389	0.6374	0.5634	0.5635

Table E.63: Stability of feature relevance for the prediction of age on a minimal input instance length of 500 characters using a stacked model on the ordered, full feature set

Feature types	Target: Age Min. No. of Characters: 500 No. of Authors: 150 Score: Avg. Spearman's ρ (ext.) N-gram ranges	Avg. Spearman's ρ (red.)	500 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)	1000 Avg. Spearman's ρ (ext.)	Avg. Spearman's ρ (red.)
DIST_CHAR	2	0.0733	0.0228	0.0064	0.0045	-0.0029	0.0157
	2-3	0.0645	-0.0005	0.0075	0.0223	-0.0016	0.0015
	2-3-4	0.0493	0.0004	0.0092	0.0073	0.0037	0.0017
	2-3-4-5	0.0451	0.0075	0.0236	0.0156	-0.0009	0.0045
DIST_CHAR_ASIS	2	0.0799	0.0056	0.0117	0.0167	0.0066	0.0003
	2-3	0.0807	0.0021	0.0149	0.0020	0.0013	0.0205
	2-3-4	0.0751	0.0083	-0.0057	-0.0054	0.0046	0.0045
	2-3-4-5	0.0688	-0.0053	0.0013	0.0040	0.0018	0.0051
DIST_CHAR_ASIS_POS	1	0.0877	0.0359	0.0129	0.0079	0.0084	0.0085
	1-2	0.0772	0.0195	0.0121	0.0085	0.0058	0.0069
	1-2-3	0.0694	0.0280	-0.0024	0.0009	0.0064	0.0082
DIST_CHAR_ASIS_POS_TAG	1	0.0568	0.0114	-0.0062	—	0.0164	0.0198
	1-2	0.0600	0.0070	0.0046	-0.0013	0.0009	0.0146
	1-2-3	0.0652	0.0193	0.0145	0.0120	0.0032	0.0120
DIST_CHAR_ASIS_POS_TAG_DEP	1	0.0593	0.0090	0.0097	0.0081	0.0046	0.0094
	1-2	0.0666	0.0221	0.0106	0.0098	0.0118	0.0109
	1-2-3	0.0577	0.0178	0.0123	0.0135	0.0120	0.0110
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA	1	0.0462	0.0048	0.0117	0.0081	0.0019	0.0065
	1-2	0.0491	0.0121	0.0052	-0.0006	0.0025	0.0052
DIST_CHAR_ASIS_POS_TAG_DEP_LEMMA_WORD	1	0.0519	0.0136	0.0101	0.0044	0.0128	0.0159
	1-2	0.0459	0.0123	0.0077	0.0107	0.0098	0.0004

BIBLIOGRAPHY

Abadie, Alberto (2018). "Statistical Non-Significance in Empirical Economics". In: December, pp. 1–25.

Agarwal, Rishabh et al. (2020). "Neural additive models: Interpretable machine learning with neural nets". In: *arXiv preprint arXiv:2004.13912*.

Aletras, Nikolaos et al. (Oct. 2016). "Predicting judicial decisions of the European court of human rights: A natural language processing perspective". In: *PeerJ Computer Science* 2016.10, e93.

Amin, Engi, Mohamed Abouelela, and Amal Soliman (2018). "The Role of Heterogeneity and the Dynamics of Voluntary Contributions to Public Goods: An Experimental and Agent-Based Simulation Analysis". In: *Journal of Artificial Societies and Social Simulation* 21.1.

Amir, Ofra, David G Rand, et al. (2012). "Economic games on the internet: The effect of $1 stakes". In: *PloS one* 7.2, e31461.

Anderhub, Vital et al. (2001). "On the Interaction of Risk and Time Preferences: An Experimental Study". In: *German Economic Review* 2.3, pp. 239–253.

Andersen, Steffen et al. (2008). "Eliciting risk and time preferences". In: *Econometrica* 76.3, pp. 583–618.

Andreoni, By James and Charles Sprenger (2012). "Risk Preferences Are Not Time Preferences". In: *The American Economic Review* 102.7, pp. 3357–3376.

Andreoni, James (1995). "Cooperation in Public-Goods Experiments: Kindness or Confusion?" In: *American Economic Review*, pp. 891–904.

Angwin, Julia et al. (May 2013). *Machine Bias. Theres software used across the country to predict future criminals. And its biased against blacks.*

Arbelaitz, Olatz et al. (2013). "An Extensive Comparative Study of Cluster Validity Indices". In: *Pattern Recognition* 46.1, pp. 243–256.

Ariely, Dan and Michael I Norton (2007). "Psychology and experimental economics: A gap in abstraction". In: *Current Directions in Psychological Science* 16.6, pp. 336–339.

Arifovic, Jasmina and John Ledyard (2012). "Individual Evolutionary Learning, Other-Regarding Preferences, and the Voluntary Contributions Mechanism". In: *Journal of Public Economics* 96.9-10, pp. 808–823.

Arnold, Jeffrey B (2018). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*.

Ash, Elliott and Daniel L. Chen (2018). "Mapping the Geometry of Law Using Document Embeddings". In: *SSRN Electronic Journal*.

Ash, Elliott, Daniel L. Chen, and Wei Lu (2018). *Motivated Reasoning in the Field: Partisanship in Precedent, Prose, Vote, and Retirement in US Circuit Courts, 1800-2013*. Tech. rep.

Athanasopoulos, Panos et al. (2010). “Perceptual shift in bilingualism: Brain potentials reveal plasticity in pre-attentive colour perception”. In: *Cognition* 116.3, pp. 437–443.

Attema, Arthur E et al. (2015). “Measuring Discounting without Measuring Utility”. In: December, pp. 1–26.

Azarbonyad, Hosein et al. (2015). “Time-aware authorship attribution for short text streams”. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 727–730.

Bail, Christopher Andrew (2016). “Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media”. In: *Proceedings of the National Academy of Sciences* 113.42, pp. 11823–11828.

Bapna, Ravi et al. (2004). “User Heterogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration”. In: *MIS Quarterly*, pp. 21–43.

Barberá, Pablo and Gonzalo Rivero (2015). “Understanding the political representativeness of Twitter users”. In: *Social Science Computer Review* 33.6, pp. 712–729.

Bartos-Höppner, Barbara (2010). *Hein Schlotterbüx aus Buxtehude: Warum in Buxtehude die Hunde mit dem Schwanz bellen*. Reissue (O. Buxtehuder Märchenbuch.

Benhabib, Jess, Alberto Bisin, and Andrew Schotter (2010). “Present-bias, quasi-hyperbolic discounting, and fixed costs”. In: *Games and Economic Behavior* 69.2, pp. 205–223.

Beriain, Iñigo De Miguel (2018). “Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling”. In: *Law, Probability and Risk* 17.1, pp. 45–53.

Berndt, Donald J. and James Clifford (1994). “Using Dynamic Time Warping to Find Patterns in Time Series.” In: *KDD workshop*. Vol. 10. 16. Seattle, WA, USA: pp. 359–370.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian (2017). “Gender stereotypes about intellectual ability emerge early and influence childrens interests”. In: *Science* 355.6323, pp. 389–391.

Boella, Guido, Luigi Di Caro, and Llio Humphreys (2011). “Using classification to support legal knowledge engineers in the Eunomos legal document management system”. In: *Fifth international workshop on Juris-informatics (JURISIN)*.

Boenninghoff, Benedikt et al. (2019). “Explainable authorship verification in social media via attention-based similarity learning”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 36–45.

Booth, Alison L and Patrick Nolen (2012). “Gender differences in risk behaviour: does nurture matter?” In: *The Economic Journal* 122.558, F56–F78.

Brennan, Tim, William Dieterich, and Beate Ehret (2009). “Evaluating the predictive validity of the COMPAS risk and needs assessment system”. In: *Criminal Justice and Behavior* 36.1, pp. 21–40.

Briley, Donnel A, Michael W Morris, and Itamar Simonson (2005). “Cultural chameleons: Biculturals, conformity motives, and decision making”. In: *Journal of Consumer Psychology* 15.4, pp. 351–362.

Brocardo, Marcelo Luiz, Issa Traore, and Isaac Woungang (2015). “Authorship verification of e-mail and tweet messages applied for continuous authentication”. In: *Journal of Computer and System Sciences* 81.8, pp. 1429–1440.

Brooks, Justin and Alexander Simpson (2012). "Find the cost of freedom: The state of wrongful conviction compensation statutes across the country and the strange legal odyssey of Timothy Atkins". In: *San Diego L. Rev.* 49, p. 627.

Buitinck, Lars et al. (2013). "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Burks, Stephen et al. (2012). "Which measures of time preference best predict outcomes: Evidence from a large-scale field experiment". In: *Journal of Economic Behavior & Organization* 84.1, pp. 308–320.

Burley, Timothy et al. (2020). "NLP Workflows for Computational Social Science: Understanding Triggers of State-Led Mass Killings". In: *Practice and Experience in Advanced Research Computing*, pp. 152–159.

Bylund, Emanuel and Panos Athanasopoulos (2017). "The Whorfian time warp: Representing duration through the language hourglass." In: *Journal of Experimental Psychology: General* 146.7, p. 911.

Cameron, Colin, Jonah Gelbach, and Douglas Miller (2006). "Robust inference with multi-way clustering". In: *NBER Working Paper* September, pp. 1–34.

Cao, Yu, Elliott Ash, and Daniel L. Chen (2018). "Automated Fact-Value Distinction in Court Opinions". In: *SSRN Electronic Journal*.

Carlson, Alyssa M (2017). "The Need for Transparency in the Age of Predictive Sentencing Algorithms". In: *Iowa Law Review* 103, pp. 303–330.

Casper, Gerhard and Richard A. Posner (June 1974). "A Study of the Supreme Court's Caseload". In: *The Journal of Legal Studies* 3.2, pp. 339–375.

Castillo, Flor A et al. (2002). "Symbolic regression in design of experiments: A case study with linearizing transformations". In: *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, pp. 1043–1047.

Caswell, Thomas A et al. (May 2021). *matplotlib/matplotlib: REL: v3.4.2*. Version v3.4.2.

Chan, Sophia and Alona Fyshe (2018). "Social and Emotional Correlates of Capitalization on Twitter". In: *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pp. 10–15.

Chandler, Seth J (2005). "The network structure of supreme court jurisprudence". In: *University of Houston Law Center* 2005-W, p. 1.

Charness, Gary and Uri Gneezy (2012). "Strong evidence for gender differences in risk taking". In: *Journal of Economic Behavior & Organization* 83.1, pp. 50–58.

Chaski, Carole E (2001). "Empirical evaluations of language-based author identification techniques". In: *Forensic Linguistics* 8, pp. 1–65.

— (2012). "Best practices and admissibility of forensic author identification". In: *JL & Pol'y* 21, p. 333.

Chaudhuri, Ananish (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature". In: *Experimental Economics* 14.1, pp. 47–83.

Chen, Daniel L, Martin Schonger, and Chris Wickens (2016). "oTree: An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.

Chen, Josie, Tai-Sen He, and Yohanes E Riyanto (2019). "The effect of language on economic behavior: Examining the causal link between future tense and time preference in the lab". In: *European Economic Review* 120, p. 103307.

Chen, Keith (Apr. 2013). "The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets". In: *American Economic Review* 103.2, pp. 690–731.

Chen, Shu-Heng, Ying-Fang Kao, and Ragupathy Venkatachalam (2017). *Computational behavioural economics*. Working Paper.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (2014). "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". In: *Journal of communication* 64.2, pp. 317–332.

Congdon, William J and Maya Shankar (2015). "The white house social & behavioral sciences team: lessons learned from year one". In: *Behavioral Science & Policy* 1.2, pp. 77–86.

Cook, Vivian et al. (June 2006). "Do bilinguals have different concepts? The case of shape and material in Japanese L2 users of English". In: *International Journal of Bilingualism* 10.2, pp. 137–152.

Costa-jussà, Marta R (2019). "An analysis of gender bias studies in natural language processing". In: *Nature Machine Intelligence* 1.11, pp. 495–496.

Crammer, Koby et al. (2006). "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7, pp. 551–585.

Crosetto, Paolo and Antonio Filippin (2013). "The "bomb" risk elicitation task". In: *Journal of Risk and Uncertainty* 47.1, pp. 31–65.

Custódio, José Eleandro and Ivandré Paraboni (2021). "Stacked authorship attribution of digital texts". In: *Expert Systems with Applications* 176, p. 114866.

Cuturi, Marco (2011). "Fast Global Alignment kernels". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 929–936.

Cuturi, Marco and Mathieu Blondel (2017). "Soft-DTW: a Differentiable Loss Function for Time Series". In: *arXiv preprint arXiv:1703.01541*.

Dahl, Östen (2000). *Tense and Aspect in the Languages of Europe*. Walter de Gruyter.

Dahl, Östen and Viveka Velupillai (2005). "The Future Tense". In: *The world atlas of language structures*. Ed. by Martin Haspelmath. Vol. 1. Oxford University Press.

Dainow, Joseph (1966). "The civil law and the common law: Some points of comparison". In: *The American Journal of Comparative Law* 15, p. 419.

Danner, Mona JE, Marie VanNostrand, and Lisa M Spruance (2015). "Risk-based pretrial release recommendation and supervision guidelines". In: *Luminosity, Inc*.

Dave, Chetan et al. (2010). "Eliciting risk preferences: When is simple better?" In: *Journal of Risk and Uncertainty* 41.3, pp. 219–243.

De Jonge, Sarah and Nenagh Kemp (2012). "Textmessage abbreviations and language skills in high school and university students". In: *Journal of Research in Reading* 35.1, pp. 49–68.

Diederich, Johannes, Timo Goeschl, and Israel Waichman (2016). "Group Size and the (In)Efficiency of Pure Public Good Provision". In: *European Economic Review* 85, pp. 272–287.

Dieterich, William, Christina Mendoza, and Tim Brennan (2016). "COMPAS risk scales: Demonstrating accuracy equity and predictive parity". In: *Northpoint Inc* 7.7.4, p. 1.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, pp. 1–15.

Djuric, Nemanja et al. (2015). "Hate speech detection with comment embeddings". In: *Proceedings of the 24th international conference on world wide web*. ACM, pp. 29–30.

Dohmen, Thomas et al. (2011). "Individual risk attitudes: Measurement, determinants, and behavioral consequences". In: *Journal of the European Economic Association* 9.3, pp. 522–550.

Dressel, Julia and Hany Farid (2018). "The accuracy, fairness, and limits of predicting recidivism". In: *Science advances* 4.1, eaao5580.

Engel, Christoph (2020). "Estimating heterogeneous reactions to experimental treatments". In: *Journal of Economic Behavior & Organization* 178, pp. 124–147.

Engel, Christoph, Sebastian Kube, and Michael Kurschilgen (2021). "Managing expectations: How selective information affects cooperation and punishment in social dilemma games". In: *Journal of Economic Behavior & Organization* 187, pp. 111–136.

Engel, Christoph and Bettina Rockenbach (2020). "What Makes Cooperation Precarious?" Working paper.

Epstein, Lee, Andrew D Martin, et al. (2007). "Ideological drift among supreme court justices: Who, when, and how important". In: *Nw. UL Rev.* 101, p. 1483.

— (2012). "Ideology and the Study of Judicial Behavior". In: *Ideology, Psychology & Law* 705.

Epstein, Lee and Jeffrey A. Segal (2005). *Advice and Consent: The Politics of Judicial Appointments*.

Everett, Daniel L (2012). *Language: The cultural tool*. Vintage.

— (Aug. 2005). "Cultural Constraints on Grammar and Cognition in Pirahã". In: *Current Anthropology* 46.4, pp. 621–646.

Fass, Tracy L et al. (2008). "The LSI-R and the COMPAS: Validation data on two risk-needs tools". In: *Criminal Justice and Behavior* 35.9, pp. 1095–1108.

Fehr, Ernst and Klaus Schmidt (2002). "Theories of Fairness and Reciprocity. Evidence and Economic Applications". In: *Advances in Economics and Econometrics. 8th World Congress*. Ed. by Mathias Dewatripont and Stephen J. Turnovsky. Cambridge: Cambridge University Press, pp. 208–257.

Ficici, Sevan G., David C. Parkes, and Avi Pfeffer (2012). "Learning and Solving Many-Player Games Through a Cluster-Based Representation". In: *arXiv preprint arXiv:1206.3253*.

Fischbacher, Urs, Simon Gächter, Nick Bardsley, et al. (2010). "SOCIAL PREFERENCES, BELIEFS, AND THE DYNAMICS OF FREE RIDING IN PUBLIC GOOD EXPERIMENTS". In: *American Economic Review* 100.1, pp. 541–556.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment". In: *Economics Letters* 71.3, pp. 397–404.

Fischman, Joshua B and David S Law (2009). "What is judicial ideology, and how should we measure it". In: *Wash. UJL & Pol'y* 29, p. 133.

Flekova, Lucie, Daniel Preoiuc-Pietro, and Lyle Ungar (2016). "Exploring stylistic variation with age and income on twitter". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 313–319.

Flores, Anthony W, Kristin Bechtel, and Christopher T Lowenkamp (2016). "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". In: *Fed. Probation* 80, p. 38.

Fong, Christian and Matthew Tyler (2018). "Machine Learning Predictions as Regression Covariates".

Forstall, Christopher and Walter Scheirer (2010). "Features from frequency: Authorship and stylistic analysis using repetitive sound". In: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*. Vol. 1. 2.

Förste, Artur Conrad (1995). *38 neue Forschungen und Quellen zur Geschichte und Ortsnamenkunde der Buxtehuder Geest*. Moisburg: Self-Published, p. 408.

Fox, John and Sanford Weisberg (2011). *An {R} Companion to Applied Regression*. Second. Thousand Oaks {CA}: Sage.

Francone, Frank D et al. (1999). "Homologous Crossover in Genetic Programming." In: *GECCO*. Citeseer, pp. 1021–1026.

Frank, Jerome and Brian H Bix (2017). *Law and the modern mind*. Routledge.

Franklin, A et al. (2008). "Lateralization of categorical perception of color changes with color term acquisition". In: *Proceedings of the National Academy of Sciences*.

Freeman, David et al. (2016). "Procedures for eliciting time preferences". In: *Journal of Economic Behavior and Organization* 126, pp. 235–242.

Gerlach, Martin et al. (2018). "A robust data-driven approach identifies four personality types across four large datasets". In: *Nature human behaviour* 2.10, pp. 735–742.

Giles, Micheal W., Virginia A. Hettinger, and Todd Peppers (Sept. 2001). "Picking federal judges: A note on policy and partisan selection agendas". In: *Political Research Quarterly* 54.3, pp. 623–641.

Ginn, Martha Humphries, Kathleen Searles, and Amanda Jones (Apr. 2015). "Vouching for the Court? How High Stakes Affect Knowledge and Support of the Supreme Court". In: *Justice System Journal* 36.2, pp. 163–179.

Gneezy, Uri and Aldo Rustichini (2004). "Gender and competition at a young age". In: *American Economic Review* 94.2, pp. 377–381.

Grgi-Hlaa, Nina, Christoph Engel, and Krishna P Gummadi (2019). "Human decision making with machine assistance: An experiment on bailing and jailing". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–25.

Grgi-Hlaa, Nina, Muhammad Bilal Zafar, et al. (2018). "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Hagan, John and Ronit Dinovitzer (1999). "Collateral consequences of imprisonment for children, communities, and prisoners". In: *Crime and justice* 26, pp. 121–162.

Halvani, Oren, Christian Winter, and Anika Pflug (2016). "Authorship verification for different languages, genres and topics". In: *Digital Investigation* 16, S33–S43.

Harris, Kamela and Rand Paul (2017). *Pretrial Integrity and Safety Act of 2017*.

Harrison, Glenn W et al. (2005). "Eliciting risk and time preferences using field experiments: Some methodological issues". In: *Field experiments in economics*. Emerald Group Publishing Limited, pp. 125–218.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hausladen, Carina I, Marcel H Schubert, and Elliott Ash (2020). "Text classification of ideological direction in judicial opinions". In: *International Review of Law and Economics* 62, p. 105903.

Heckathorn, Douglas D (1989). "Collective action and the second-order free-rider problem". In: *Rationality and Society* 1.1, pp. 78–100.

Hlavac, Marek (2016). "ExtremeBounds : Extreme Bounds Analysis in R". In: *Journal of Statistical Software* 72.9.

— (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia.

Horton, John J, David G Rand, and Richard J Zeckhauser (2011). "The online laboratory: Conducting experiments in a real labor market". In: *Experimental economics* 14.3, pp. 399–425.

Houser, Daniel, Kevin McCabe, Vernon Smith, et al. (2004). "Cultural group selection, co evolutionary processes and large-scale cooperation (by Joseph Henrich)". In: *Journal of economic behavior & organization* 53.1, pp. 85–88.

Hovy, Dirk and Anders Søgaard (2015). "Tagging performance correlates with author age". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 483–488.

Huang, Wenjing, Rui Su, and Mizuho Iwaihara (2020). "Contribution of improved character embedding and latent posting styles to authorship attribution of short texts". In: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, pp. 261–269.

InternetLiveStats (2019). *Twitter Usage Statistics*. Last Accessed: 20-08-2019.

Isaac, R. Mark and James M. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism". In: *Quarterly Journal of Economics* 103.1, pp. 179–199.

Jackson, Eugenie and Christina Mendoza (2020). "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not". In: *2.1* 2.1, pp. 1–15.

Johnson, Susan W., Donald R. Songer, and Nadia A. Jilani (July 2011). "Judge gender, critical mass, and decision making in the appellate courts of Canada". In: *Journal of Women, Politics and Policy* 32.3, pp. 237–260.

Jung, Jongbin, Sharad Goel, Jennifer Skeem, et al. (2020). "The limits of human predictions of recidivism". In: *Science Advances* 6.7, eaaz0652.

Kassow, Benjamin, Donald R. Songer, and Michael P. Fix (2012). "The Influence of Precedent on State Supreme Courts". In: *Political Research Quarterly* 65.2, pp. 372–384.

Keelj, Vlado et al. (2003). "N-gram-based author profiles for authorship attribution". In: *Proceedings of the conference pacific association for computational linguistics, PACLING*. Vol. 3. sn, pp. 255–264.

Kim, Yoon (2014). *Convolutional neural networks for sentence classification*. Tech. rep.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (2011). "Authorship attribution in the wild". In: *Language Resources and Evaluation* 45.1, pp. 83–94.

Kosfeld, Michael, Akira Okada, and Arno Riedl (2009). "Institution Formation in Public Goods Games". In: *American Economic Review* 99.4, pp. 1335–55.

Kotanchek, Mark, Guido Smits, and Arthur Kordon (2003). "Industrial strength genetic programming". In: *Genetic programming theory and practice*. Springer, pp. 239–255.

Kreps, David M et al. (1982). "Rational cooperation in the finitely repeated prisoners' dilemma". In: *Journal of Economic Theory* 27.2, pp. 245–252.

Lahey, Benjamin B et al. (2000). "Age and gender differences in oppositional behavior and conduct problems: a cross-sectional household study of middle childhood and adolescence." In: *Journal of abnormal psychology* 109.3, p. 488.

Landes, William M. and Richard A. Posner (2009). "Rational Judicial Behavior : A Statistical Study". In: *Journal of Legal Analysis* 1.2, pp. 775–831.

Larson, Jeff and Mattu, Surya and Kirchner, Lauren and Angwin, Julia (n.d.). *How We Analyzed the COMPAS Recidivism Algorithm*. Accessed: 2021-07-11.

Laub, Zachary (2019). *Hate Speech on Social Media: Global Comparisons*. Last Accessed: 20-08-2019.

Lauderdale, Benjamin E and Tom S Clark (2014). "Scaling politically meaningful dimensions using texts and votes". In: *American Journal of Political Science* 58.3, pp. 754–771.

Lauderdale, Benjamin E and Alexander Herzog (2016). "Measuring Political Positions from Legislative Speech". In: *Political Analysis*, pp. 1–21.

Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data". In: *The American Political Science Review* 97.2, pp. 311–331.

Leamer, Edward E. (1985). "Sensitivity Analyses Would Help". In: *The American Economic Review* 75.3, pp. 308–313.

Ledyard, John O (1995). "Public Goods: A Survey of Experimental Research". In: *Handbook of Experimental Economics*. Ed. by John Kagel and Al Roth. Princeton: Princeton University Press, pp. 111–194.

Lessig, Lawrence (1999). *Code: And other laws of cyberspace*. ReadHowYouWant. com.

Levitt, Steven D and John A List (2007). "What do laboratory experiments measuring social preferences reveal about the real world?" In: *Journal of Economic perspectives* 21.2, pp. 153–174.

Li, King King (Aug. 2017). "How does language affect decision-making in social interactions and decision biases?" In: *Journal of Economic Psychology* 61, pp. 15–28.

Liao, T. Warren (2005). "Clustering of Time Series Data – a Survey". In: *Pattern Recognition* 38.11, pp. 1857–1874.

Lindquist, Kristen A. et al. (2006). "Language and the perception of emotion." In: *Emotion* 6.1, pp. 125–138.

Lu, Yixin et al. (2016). "Exploring Bidder Heterogeneity in Multichannel Sequential B2B Auctions". In: *MIS Quarterly* 40.3, pp. 645–662.

Lucas, Pablo, Angela de Oliveira, and Sheheryar Banuri (2012). "The Effects of Group Composition and Social Preference Heterogeneity in a Public Goods Game: An Agent-Based Simulation". In: *Journal of Artificial Societies and Social Simulation* 17.3, pp. 148–174.

Luna, David, Torsten Ringberg, and Laura A Peracchio (2008). "One individual, two identities: Frame switching among biculturals". In: *Journal of Consumer Research* 35.2, pp. 279–293.

Majid, Asifa et al. (2004). "Can language restructure cognition? The case for space". In: *Trends in Cognitive Sciences* 8.3, pp. 108–114.

Manski, Charles F et al. (2020). "Judicial and Clinical Decision-Making under Uncertainty". In: *Journal of Institutional and Theoretical Economics (JITE)* 176.1, pp. 33–43.

Mao, Andrew et al. (2017). "Resilient cooperators stabilize long-run cooperation in the finitely repeated Prisoners Dilemma". In: *Nature communications* 8.1, pp. 1–10.

Martin, Andrew D. and Kevin M. Quinn (2001). "The Dimensions of Supreme Court Decision Making : Again Revisiting The Judicial Mind". In: pp. 1–37.

— (2002). "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999". In: *Political Analysis* 10.2, pp. 134–153.

Martin, Andrew D., Kevin M. Quinn, and Lee Epstein (2004). *The median justice on the united states supreme court*. Tech. rep., p. 1275.

Masood, Ali S. and Donald R. Songer (2013). "Reevaluating the Implications of Decision-Making Models". In: *Journal of Law and Courts* 1.2, pp. 363–389.

McFadden, Daniel (1973). "Conditional logit analysis of qualitative choice behavior". In.

Michael, Eberhard (2018). *Deutsche MärchenstraSSe. Ein Reise- und Lesebuch mit Märchen, Sagen und Legenden*. Ed. by CW Niemeyer Buchverlage. 2nd. Hameln.

Monti-Belkaoui, Janice and Ahmed Belkaoui (1983). "Bilingualism and the perception of professional concepts". In: *Journal of Psycholinguistic Research* 12.2, pp. 111–127.

Narayanan, Arvind et al. (2012). "On the feasibility of internet-scale author identification". In: *2012 IEEE Symposium on Security and Privacy*. IEEE, pp. 300–314.

Neal, Tempestt et al. (2017). "Surveying Stylometry Techniques and Applications". In: *ACM Computing Surveys* 50.6, pp. 1–36.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.

Nikiforakis, Nikos and Hans-Theo Normann (2008). "A Comparative Statics Analysis of Punishment in Public-Good Experiments". In: *Experimental Economics* 11.4, pp. 358–369.

Nishi, Andrea (2019). "Privatizing Sentencing". In: *Columbia Law Review* 119.6, pp. 1671–1710.

Noorian, Farzad, Anthony M. de Silva, and Philip H. W. Leong (2016). "gramEvol: Grammatical Evolution in R". In: *Journal of Statistical Software* 71.1, pp. 1–26.

Northpointe (2015). *Practitioners Guide to COMPAS*. Accessed: 2021-04-30.

PAN (2019). *Celebrity Profiling*. Last Accessed: 20-08-2019.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis (2010). "Running experiments on amazon mechanical turk". In.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035.

Pavlick, Ellie et al. (2016). "The Gun Violence Database: A new task and data set for NLP". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1018–1024.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Peng, Fuchun et al. (2003). "Language independent authorship attribution using character level language models". In: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 267–274.

Petitjean, François, Alain Ketterlin, and Pierre Gançarski (2011). "A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering". In: *Pattern Recognition* 44.3, pp. 678–693.

Platt, John C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in large margin classifiers* 10.3, pp. 61–74.

PreuSSler, Ottfried (1962). *Der Räuber Hotzenplotz*. 45th ed. Thienemann Verlag in der Thienemann-Esslinger Verlag GmbH.

Pritchett, Lant and Justin Sandefur (2014). "Context matters for size: Why external validity claims and development practice do not mix". In: *Journal of Globalization and Development* 4.2, pp. 161–197.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Randazzo, Kirk A., Richard W. Waterman, and Michael P. Fix (2010). "State Supreme Courts and the Effects of Statutory Constraint". In: *Political Research Quarterly* 64.4, pp. 779–789.

Regier, Terry and Paul Kay (Oct. 2009). "Language, thought, and color: Whorf was half right". In: *Trends in Cognitive Sciences* 13.10, pp. 439–446.

Reid, Rebecca and Kirk A. Randazzo (July 2016). "Statutory Language and the Separation of Powers". In: *Justice System Journal* 37.3, pp. 246–258.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 1135–1144.

— (2016b). "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 1135–1144.

Rifkin, Ryan and Ross Lippert (2007). *Notes on Regularized Least Squares*. Tech. rep. Cambridge: Massachusetts Institute of Technology, pp. 1–10.

Rocha, Anderson et al. (2017). "Authorship Attribution for Social Media Forensics". In: *IEEE Transactions on Information Forensics and Security* 12.1, pp. 5–33.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA.

Rudin, Cynthia (Nov. 2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215.

Rudin, Cynthia, Caroline Wang, and Beau Coker (2020). "Broader Issues Surrounding Model Transparency in Criminal Justice Risk Scoring". In: *Harvard Data Science Review* 2.1, pp. 1–16.

Sackett, David L et al. (1996). *Evidence based medicine: what it is and what it isn't*.

Sage, Manuel et al. (2020). "Investigating the Influence of Selected Linguistic Features on Authorship Attribution using German News Articles." In: *SwissText/KONVENS*.

Sala-i-Martin, Xavier X. (1997). "I just ran four million regressions".

Samek, Wojciech et al. (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

Sampson, Robert J and John H Laub (1992). "Crime and deviance in the life course". In: *Annual review of sociology* 18.1, pp. 63–84.

Sanchez-Perez, Miguel A et al. (2017). "Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 145–151.

Sardá-Espinosa, Alexis (2017). "Comparing Time Series Clustering Algorithms in R using the dtwclust Package". In: *R Package Vignette* 12, p. 41.

Schmidt, Jan-Hinrik (2014). "Twitter and the rise of personal publics". In: *Twitter and society*, pp. 3–14.

Schmidt, Mark, Nicolas Le Roux, and Francis Bach (2017). "Minimizing finite sums with the stochastic average gradient". In: *Mathematical Programming* 162.1-2, pp. 83–112.

Segal, Jeffrey A et al. (1995). "Ideological values and the votes of US Supreme Court justices revisited". In: *The Journal of Politics* 57.3, pp. 812–823.

Segal, Jeffrey A. and Albert D. Cover (1989). "Ideological values and the votes of US Supreme Court justices". In: *American Political Science Review* 83.2, pp. 557–565.

Shrestha, Prasha et al. (2017). "Convolutional neural networks for authorship attribution of short texts". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 669–674.

Sidorov, Grigori et al. (2014). "Syntactic n-grams as machine learning features for natural language processing". In: *Expert Systems with Applications* 41.3, pp. 853–860.

Slapin, Jonathan B. and Sven-Oliver Proksch (2008). "A Scaling Model for Estimating Time-Series Part Positions from Texts". In: *American Journal of Political Science* 52.3, pp. 705–722.

Songer, Donald R. (1993). "The United States Court of Appeals Database - Documentation for Phase I".

Spitters, Martijn et al. (2016). “Authorship Analysis on Dark Marketplace Forums”. In: *Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015*, pp. 1–8.

Stamatatos, Efstathios (2009). “A survey of modern authorship attribution methods”. In: *Journal of the American Society for information Science and Technology* 60.3, pp. 538–556.

— (2013). “On the robustness of authorship attribution based on character n-gram features”. In: *Journal of Law and Policy* 21.2, pp. 421–439.

— (2017). “Authorship attribution using text distortion”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1138–1149.

Stamatatos, Efstathios et al. (2015). “Overview of the pan/clef 2015 evaluation lab”. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pp. 518–538.

Sturm, H. P. and C. Herman Pritchett (2006). “The Roosevelt Court, a Study in Judicial Politics and Values, 1937-1947”. In: *The Western Political Quarterly* 2.3, p. 465.

Stuttgarter Nachrichten (2017). *Bitte, Facebook, Hilf uns*. URL: `https://www.stuttgarter-nachrichten.de/inhalt.polizei-als-bittsteller-bei-straftaten-bitte-facebook-hilf-uns.a3a743a3-d41e-479c-aba7-d5bc4da25286.html` (visited on 07/22/2021).

Suen, Ching Y. (1979). “N-gram statistics for natural language understanding and text processing”. In: *IEEE transactions on pattern analysis and machine intelligence* 2, pp. 164–172.

Sulea, Octavia Maria et al. (2017). “Exploring the use of text classification in the legal domain”. In: *CEUR Workshop Proceedings* 2143.

Sutter, Matthias, Silvia Angerer, et al. (Sept. 2015). “The Effect of Language on Economic Behavior: Experimental Evidence from Children’s Intertemporal”. Working Paper.

Sutter, Matthias and Daniela Glätzle-Rützler (2014). “Gender differences in the willingness to compete emerge early in life and persist”. In: *Management Science* 61.10, pp. 2339–2354.

Theophilo, Antonio, Luis A. M. Pereira, and Anderson Rocha (2019). “A Needle in a Haystack? Harnessing Onomatopoeia and User-specific Stylometrics for Authorship Attribution of Micro-messages”. In: pp. 2692–2696.

Thompson, Bill, Seán G Roberts, and Gary Lupyan (2020). “Cultural influences on word meanings revealed through large-scale semantic alignment”. In: *Nature Human Behaviour*, pp. 1–10.

Udrescu, Silviu-Marian and Max Tegmark (2020). “AI Feynman: A physics-inspired method for symbolic regression”. In: *Science Advances* 6.16, eaay2631.

Undavia, Samir, Adam Meyers, and John Ortega (2018). “A Comparative Study of Classifying Legal Documents with Neural Networks”. In: *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems* 15.October, pp. 515–522.

Van Nes, Fenna et al. (2010). “Language differences in qualitative research: is meaning lost in translation?” In: *European journal of ageing* 7.4, pp. 313–316.

Varian, Hal R (2014). “Big data: New tricks for econometrics”. In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Varol, Onur et al. (2017). “Online human-bot interactions: Detection, estimation, and characterization”. In: *Eleventh international AAAI conference on web and social media*.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008.

Verhoeven, Ben, Walter Daelemans, and Barbara Plank (2016). “Twisty: a multilingual twitter stylometry corpus for gender and personality profiling”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1632–1637.

Vorobeychik, Yevgeniy, Michael P. Wellman, and Satinder Singh (2007). “Learning Payoff Functions in Infinite Games”. In: *Machine Learning* 67.1-2, pp. 145–168.

Wagner, Gert G, Joachim R Frick, and Jürgen Schupp (2007). “The German Socio-Economic Panel study (SOEP)-evolution, scope and enhancements”. In.

Waskom, Michael L. (2021). “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60, p. 3021.

Watson, James; Björn; Holm, and Brandt; Heike (1992). *The Noisy Ducks of Buxtehude*. Buxtehude: Vlg. an der Este.

Western, Bruce, Jeffrey R Kling, and David F Weiman (2001). “The labor market consequences of incarceration”. In: *Crime & delinquency* 47.3, pp. 410–427.

Western, Bruce, Leonard Lopoo, and Sara McLanahan (2004). “Incarceration and the bonds among parents in fragile families”. In: *Imprisoning America: The social effects of mass incarceration*, pp. 21–45.

Wickham, Hadley (2011). “The Split-Apply-Combine Strategy for Data Analysis”. In: *Journal of Statistical Software* 40.1, pp. 1–29.

— (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

— (2018). *stringr: Simple, Consistent Wrappers for Common String Operations*.

Wiegmann, Matti, Benno Stein, and Martin Potthast (2019). “Overview of the Celebrity Profiling Task at PAN 2019.” In: *CLEF (Working Notes)*.

Winawer, Jonathan et al. (2007). “Russian blues reveal effects of language on color discrimination”. In: *Proceedings of the National Academy of Sciences* 104.19, pp. 7780–7785.

Yamagishi, Toshio (1986). “The provision of a sanctioning system as a public good”. In: *Journal of Personality and Social Psychology* 51.1, pp. 110–116.

Zeileis, Achim (2004). “Econometric Computing with HC and HAC Covariance Matrix Estimators”. In: *Journal of Statistical Software* 11.10, pp. 1–17.

— (2006). “Object-Oriented Computation of Sandwich Estimators”. In: *Journal of Statistical Software* 16.9, pp. 1–16.

Zeileis, Achim and Torsten Hothorn (2002). “Diagnostic Checking in Regression Relationships”. In: *R News* 2.3, pp. 7–10.

Zelmer, Jennifer (2003). “Linear Public Goods Experiments: A Meta-Analysis”. In: *Experimental Economics* 6.3, pp. 299–310.

Zhang, Tong (2004). “Solving large scale linear prediction problems using stochastic gradient descent algorithms”. In: *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 919–926.

Zufall, Frederike et al. (2020). “Operationalizing the legal concept of’Incitement to Hatred’as an NLP task”. In: *arXiv preprint arXiv:2004.03422*.

Marcel H. Schubert

Personal Information

First Name / Last Name	Marcel H. Schubert
Address	Aalen, Germany
Mail	schubert[@]wiso.uni-koeln.de
Nationality	German

Fields of Competence

Expertise in Data Science and Machine Learning

- Proficient with current libraries front- and back-ends (i.e., PyTorch, TensorFlow etc.)
- Expertise with NLP as well as ML models used for NLP tasks
- Experienced with non-linear CNNs, NNs as well as linear models

Expertise in Economics

- Extensive knowledge on statistics, econometrics and their application
- Skilled in the design and the conducting of laboratory experiments

Expertise in IT Security

- Experience in analyzing privacy and the impact of machine learning models on anonymity
- Knowledge about security of computer systems and the exploitation of their weaknesses

Academic Background

09/17 - today	**PhD in the field of Human Behavior and Machine Learning, University of Cologne** Focus: Application of machine learning methods to behavioral data from behavioral economics & Social Science
10/17 - 11/19	**Master of Science Cyber Security, Lancaster University** - Focus: Security of modern systems and their vulnerabilities - Master Thesis: Forensic Authorship Analysis and Feature Stability
10715 - 09/17	**Bachelor of Science in Computer Science, University of Bonn** - Focus: Machine Learning, IT Security - Bachelor Thesis: Privacy Impact Assessment
10/15 - 09/17	**Master of Science in Economics, University of Bonn** - Focus: Microeconomics and Behavioral Economics - Master Thesis: Language and Behaviour: Does Framing through Language affect Judgement and Choice in the Domains of Risk and Time?
10/12 - 02/15	**Bachelor of Science in Economics, LMU München** - Focus: Microeconomics and Behavioral economics - Bachelor thesis: Do fMRI studies fundamentally improve our understanding of decisions involving risk, time delay or others?

Professional Experience & Research Stays

03/19 - 05/2019	Visiting Research Fellow, Center for Law, Economics, and Data Science, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland
09/17 - today	Research Fellow IMPRS, Max-Planck Institute for Research on Collective Goods, Bonn
09/15 - 09/17	Research Assistant, Max-Planck Institute for Research on Collective Goods, Bonn
10715 - 09/17	Bachelor of Science in Computer Science, University of Bonn
05/15 - 07/15	Intern, Helbling Business Advisors, Stuttgart

Publications

Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text Classification of ideological direction in judicial opinions. International Review of Law and Economics, 62, 105903.

Scholarships & Funding

04/19	IPAK Travel Grant, DAAD & University of Cologne, Cologne, Germany, $500
04/19	IPAK Travel Grant, DAAD & University of Cologne, Cologne, Germany, € 1500
07/12 - 09/17	Konrad-Adenauer Scholarship for Gifted Students ca.€ 18,000

Conferences & Talks

03/21	Engel, C., Hausladen, C. I., Schubert, M. H. (Working Paper). Charting the Type Space: The Case of Linear Public-Good Experiments. Digital Session (Covid).
04/19	Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text Classification of ideological direction in judicial opinions. PELS Replication Conference, Claremont McKenna College, Claremont, California.
02/19	Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text Classification of ideological direction in judicial opinions. IMPRS Uncertainty Thesis Workshop, Wittenberg, Germany.
02/18	Albrech, F., Schubert, M. H. (Working Paper). The Effect of Grammatical Variation on Economic Behavior: Varying Future Time References within the German Language. IMPRS Uncertainty Thesis Workshop, Schloss Ringberg, Germany.

Köln, den August 19, 2021_____

Marcel H. Schubert

SUPERVISORS

Prof. Dr. Martin Fochmann
martin.fochmann[at]fu-berlin.de
Department of Finance, Accounting, Controlling and Taxation
Thielallee 73, 14195 Berlin

Prof. Dr. Dr. h.c. Christoph Engel
engel[at]coll.mpg.de
Director of the Max Planck Institute for Research on Collective Goods
Kurt-Schumacher-Straße 10, 53113 Bonn

Eidesstattliche Versicherung

"Hiermit versichere ich an Eides Statt, dass ich die vorgelegte Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Aussagen, Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich (zutreffendes bitte unterstreichen) geholfen:

Weitere Personen neben den in der Einleitung der Dissertation aufgeführten Koautorinnen und Koautoren waren an der inhaltlich-materiellen Erstellung der vorliegenden Dissertation nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Dissertation wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe"

Köln, den August 19, 2021_____

Marcel H. Schubert