Tracing Life Events and Their Polarities in 18th Century Biographical Sources with Transformer Models

Von der Philosophischen Fakultät der Universität zu Köln angenommene
Inaugural-Dissertation
zur Erlangung des Doktorgrades

Im Fach Digital Humanities

vorgelegt von

Patrick D. Brookshire

Köln, 2026

Erklärung

Ich versichere eidesstattlich, dass ich die von mir vorgelegte Dissertation selbständig und ohne unzulässige Hilfe angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Dissertation einschließlich Tabellen, Karten und Abbildungen, die anderen fremden Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Hochschule zur Prüfung vorgelegen hat; dass sie, gegebenenfalls abgesehen von einer durch die Vorsitzende oder den Vorsitzenden des Promotionsausschusses nach Rücksprache mit der Betreuerin beziehungsweise dem Betreuer vorab genehmigten Teilpublikation, noch nicht veröffentlicht worden ist sowie dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen in §§ 20 und 21 der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Nils Reiter betreut worden.

Acknowledgements

A Ph.D. project is no sprint but a marathon which is why I am very grateful for all the support I got on the way. First and foremost, I would like to thank my supervisors, Nils Reiter and Frederik Elwert, for the many critical but always very fruitful discussions over the past years despite the distance between the three working places. Without their very much treasured feedback, guidance, and support, this work would definitely not have been possible. It really was a pleasure to get their perspectives – especially because both were always open-minded for potential new directions no matter if these proved worth investigating further or needed to be dismissed in the light of other new insights. In a similar vein, it was inspiring how they seemingly knew which conference would be the best to get even more valuable feedback and/or new ideas on how to move on.

In addition, I want to thank Elisa Cugliana for becoming the third examiner in the ending phase of this PhD project. Similarly, I like to thank Wolfgang Breul and Jonas Kuhn who both helped shaping the very first ideas of what would eventually become this work. In this sense, I thank Katie Faull as well – not only for an online meeting during the preparatory phase but also for enabling this work with the digitization efforts of hers and her colleagues over the past decades.

Moreover, I thank the Spinfo group including Melanie Andresen, Johanna Binnewitt, Michael Göggelmann, Jürgen Hermes, Lea Krumbach, Ziyue Liu, Judith Nester, Janis Pagel, and Axel Pichler very much for all the memorable colloquia, paper discussions, and most notably retreats with joint chores, guided tours, (group) presentations, and invaluable brainstorming sessions.

I also want to express my gratitude to Andrea Rapp, Claudius Geisler, Roland Kehrein, and many others from the Academy of Sciences and Literature | Mainz for enabling conference participations as well as further qualifications that had a lasting impact not only on this work. Considering my working place, many words of thanks go to all my colleagues from the Digital Academy represented by Hans-Werner Bartz, Aline Deicke, and Thorsten Schrade as this work would not have been put to an end without their support. In particular, this also applies to Alexandra Büttner, Chantal Gärtner, Jonathan D. Geiger, Philipp Hegel, Georg Hertkorn, Julian Jarosch, Jakob Jünger, Michael Haft, Paul Herdt, Dominik Kasper, Thomas Kollatz, Andreas Kuczera, Maximilian Michel, Lea Müller-Dannhausen, Joshua Neumann, Sarah Pittroff, Berenike Rensinghoff, Zahia Schlott, Annabella Schmitz, Clara Seibold, Martin Sievers, Linnaea Söhn, Jonatan Jalle Steller, Markus Studer, Elena Suárez Cronauer, and Aramis for every once in a while having "a few minutes for yet another finding". Furthermore, I thank Gabriel Belinga Belinga, Aglaia Schieke, and Gabriele Schweickhardt for showing me new perspectives on science communication to address broader audiences.

Last not least, I want to extent my thanks, from the bottom of my heart, to Susanne, Hiltrud, Hans, Johanna, Arthur, Susanne, Ralf, and Jacqueline for all their unconditional emotional support not only during these past years.

Patrick D. Brookshire

Contents

List of Figures

List of Tables

Abstract

This thesis shows how life events and their polarities can be traced in (historical) biographies based on text classification tasks. Biographies are valuable sources not only for sociological but also historical research as they deal with typical life courses which are interrelated with broader sociocultural peculiarities worth investigating. In this context, the 18th century is noteworthy due to the emergence of many biographical subgenres. In addition, many text collections have been handed down to this day and some have already become computationally assessable.

Methodologically, this work illustrates that manual annotations with (i) sentiment, (ii) life stage, and (iii) life event domain labels can be scaled up to larger-scale empirical research. To this end, promising systems discussed in related work are compared to each other in classification experiments in order to assess their applicability. In the case of the established sentiment classification task, transformer models fine-tuned with data from the target domain performed best and yielded state-of-the-art performances. Similar but slightly worse results were achieved by Large Language Models (LLMs) hosted on institutional servers. By contrast, off-the-shelf transformer models trained on modern-day sentiment datasets and zero-shot approaches other than the evaluated LLMs yielded very poor results. The same is true for popular dictionary-based approaches. Notably, Support Vector Machines (SVMs) – another supervised learning approach – also outperformed most systems but fine-tuned transformers and LLMs. Hence, one can conclude that domain differences are a major limitation of systems adapted to modern-day texts.

Similar conclusions can be drawn from the other two classification tasks which used novel categorization schemes derived from sociological and psychological research. More specifically, transformer models again outperformed all other evaluated systems by a margin after fine-tuning. SVMs yielded slightly lower performances even when optimizing the underlying vectorizer. Conversely, most zero-shot approaches performed very poorly. Still, LLMs yielded at least performances similar to some SVM settings. Most of these results are in line with findings from related work. However, they imply that deep learning models, such as transformers, outperform more transparent approaches like dictionaries and SVMs in these tasks.

Because better performing systems are preferable for downstream analyses, explainability models were applied to mitigate transparency issues. The latter suggested that the most performant classifier has implicitly learned to (i) recognize relevant textual features and (ii) follow annotation guidelines when labeling ambiguous cases.

This actually applied to a lesser degree to the life stage classifier which is arguably the most difficult one of the three but also an apt addition. For instance, both manual and automatic annotations suggested that life narratives with a female subject focused more on earlier life stages whereas later ones were described in more detail in the case of the opposite gender. Moreover, said annotations also enable a rescaling from the text course to the underlying life course. This change of perspective offered further analytical options for uncovering less overt patterns, such as the similar depiction of later life stages in different biographical subgenres. In this regard, it is worth mentioning that (sub)genre differences in sentiment polarity were more common on the narrative side while topical differences concerned both text and life courses. These exemplary findings underline already that (i) sentiment polarity is a relevant category when analyzing these sources and (ii) a more holistic analysis requires task combinations. More such findings will be discussed in this thesis and related to existing biographical research. In this sense, this work builds up on related research endeavors while proposing a methodological approach and describing notable patterns identified with it so that it in turn opens up many future directions for digital biographical research.

Deutsche Zusammenfassung

Diese Arbeit zeigt, wie sich mithilfe von Textklassifikationsaufgaben Lebensereignisse und deren Polarität in (historischen) Biographien aufdecken lassen. Biographien sind nicht nur wertvolle Quellen für soziologische Forschung sondern auch für historische, da sie typische Lebensverläufe nachzeichnen, die auf größere soziokulturelle Hintergründe bezogen werden können. In diesem Zusammenhang ist das 18. Jahrhundert hervorzuheben, in dem viele biographische Textformen entstanden sind. Desweiteren liegen viele Textsammlungen aus dieser Zeit vor, von denen einige auch so digitalisiert wurden, dass sie sich mit digitalen Methoden analysieren lassen.

Aus methodischer Sicht zeigt die Arbeit, dass manuelle Annotationen bezogen auf (i) Polarität, (ii) Lebensabschnitte und (iii) Lebensereignisbereiche für größer angelegte empirische Untersuchungen hochskaliert werden können. Zu diesem Zweck werden als erstes verschiedene in der Literatur diskutierte Ansätze im Sinne von Klassifikationsexperimenten miteinander verglichen, um abzuschätzen, wie gut sie sich für die jeweiligen Aufgaben eignen. Im Fall der etablierten Polaritätsklassifizierung haben sich Transformermodelle als am vielversprechendsten herausgestellt, wenn ein Fine-tuning mit Daten aus der Zieldomäne erfolgte. Große Sprachmodelle (LLMs), die auf institutionellen Servern lokal gehostet wurden, waren ähnlich gut, wenn auch etwas schlechter. Im Gegensatz dazu lieferten sowohl vorliegende Transformermodelle, die mit zeitgenössischen Polaritätsdaten trainiert wurden, als auch Zero-shot-Ansätze, die keine LLMs sind, vergleichsweise schlechte Ergebnisse. Dasselbe gilt für die häufig genutzten Abgleiche mit Werten in Polaritätswörterbüchern. Interessanterweise lieferten Support-Vektor-Maschinen (SVMs), die wie Transformermodelle ein Verfahren des Überwachten Machinellen Lernens sind, bessere Ergebnisse. Daher ist davon auszugehen, dass der Domänenwechsel ein größeres Problem für Systeme ist, die mit zeitgenössischen Daten trainiert wurden.

Ähnliche Schlüsse sind auch in Bezug auf die übrigen zwei Klassifikationsaufgaben möglich, die Kategorisierungen aus der Soziologie und Psychologie auf neue Weise operationalisieren. Bei diesen Aufgaben lieferten Transformermodelle erneut die mit Abstand besten Ergebnisse, dicht gefolgt von SVMs mit optimierter Vektorisierung. Umgekehrt lieferten Zero-shot-Ansätze sehr schlechte Ergebnisse, selbst wenn LLMs zumindest an einige SVM-Varianten heranreichten. Die unterschiedlichen Leistungsstärken passen größtenteils zur Forschungsliteratur. Sie implizieren jedoch auch, dass Deep-Learning-Verfahren, zu denen Transformermodelle zählen, die Klassifikationsaufgaben dieser Arbeit besser erfüllen als nachvollziehbarere Operationalisierungen wie Wörterbuchabgleiche oder SVMs.

Da besser geeignete Systeme für nachgelagerte Analyseschritte vorzuziehen sind, wurden Erklärungssysteme genutzt, um die Transparenz zu erhöhen. Diese deuteten an, dass das am besten geeignete Klassifikationssystem gelernt hat, (i) relevante Textmerkmale zu erkennen, und (ii) den Annotationsrichtlinien beim Auflösung ambiger Textstellen zu folgen.

Beides trifft nur eingeschränkt auf die Lebensabschnittsklassifikation zu, was daran liegen dürfte, dass diese die mutmaßlich schwierigste der drei Aufgaben ist. Nichtsdestotrotz ist sie eine passende Ergänzung. So verdeutlichten etwa sowohl die entsprechenden manuellen als auch automatischen Annotationen, dass Frauenbiographien einen stärkeren Fokus auf frühere Lebensabschnitte legten, während bei Männern spätere detaillierter dargestellt wurden. Darüber hinaus lassen sich die annotierten Lebensabschnitte dazu nutzen, um Verteilungsmuster nicht nur auf den Textverlauf, sondern auch den zugrundeliegenden Lebensverlauf zu beziehen. Daraus ergaben sich weitere Analysemöglichkeiten, mit denen weniger offensichtliche Ähnlichkeiten wie etwa die Darstellung späterer Abschnitte in verschiedenen biographischen Textformen aufgedeckt

werden konnten. In diesem Zusammenhang ist hervorzuheben, dass Polaritätsunterschiede zwischen verschiedenen (Teil-)Genres fast ausschließlich den Textverlauf betrafen, während thematische Schwerpunktsetzungen sich gleichermaßen in Text- und Lebensverläufen manifestierten. Diese Beispiele für inhaltliche Erkenntnisse zeigen bereits, dass (i) Polarität eine relevante Analysekategorie ist, und (ii) eine ganzheitlichere Analyse nur durch die Kombination verschiedener Klassifikationsaufgaben möglich ist. Weitere Erkenntnisse werden im Rahmen dieser Arbeit auf Ergebnisse ähnlicher Forschungsvorhaben bezogen. Auf diese Weise baut diese Arbeit auf anderen auf, während sie gleichzeitig einen Methodenbeitrag sowie erste inhaltliche Erkenntnisse liefert, mit denen viele Ansätze für künftige digitale Analysen biographischer Quellen eröffnet werden.

Chapter 1

Introduction

Biographical sources play a major role in the social sciences and have gained an increasing amount of attention in (digital) historical research over the past few decades. This is partly due to the rise of digital methods which enable quantitative analyses of larger text corpora. In this way, views of "common" individuals have become assessable and can be compared to canonical views of the past which are smaller in scale. From a literary standpoint, these sources are also worth investigating because they "may employ formal strategies from drama, poetry, essay, and fiction" and therefore offer "a richness unparalleled by any other form of writing" in the words of Gunzenhauser (2001, 77). This is especially true for the 18th century which is not only characterized by the emergence of many biographical subgenres but also the "'discovery' of childhood" as Taylor (2016, 3) put it. As a result, biographies from this time frame may cover the whole life course from childhood to old-age so that holistic analyses are possible.

Despite these noteworthy peculiarities, current digital biographical research rather focuses on digitization efforts and network analyses instead of computational content and/or stylistic analyses. More specifically, only a few studies have so far tried to identify general patterns in the depiction of the life course in general and the life events forming it in particular in 18th century biographies. In order to fill this research gap and to enable further studies, this work proposes a methodological approach based on classification tasks on the one hand, and illustrates how it can be applied to answer biographical research questions on the other.

The proposed methodology follows a scalable reading approach where manual annotations are scaled up to larger samples. In order to cover multiple perspectives, this work deals with three types of annotations: As a first annotation layer, selected biographical texts were labeled with ternary sentiment polarity. This enables a sentiment analysis which has become a widespread method in digital humanities.[1] Life stages, i.e. age-based subdivisions of the life course, are the second annotation layer and closer tied to the analysis of life narratives than the first one. As a matter of fact, said second task is a novel one – at least to the best of my knowledge. It encompasses annotating at which life stage the respective biographical subject experienced an episode of her or his life as described in a biography. Such annotations can be used to verify common conceptions, like the one that biographies follow the life course in chronological order. When combined with the first task, they also enable an assessment of whether certain life stages, such as *childhood*, are depicted more positively than others. However, it should be noted in this context that such a potential polarity difference relates only to the textual depiction. This is especially important when analyzing childhood experiences since even autobiographical accounts are typically formulated later so that they may be distorted to some extent (Clausen, 1998, 196). Nevertheless, they may still follow patterns worth investigating. The same is true for the final annotation layer which deals with life event domains. While the exact operationalization is again novel, the underlying goal of identifying major life events is not. For instance, Dennis-Henderson et al. (2020) and Haehner et al. (2024) both analyzed the polarity of life events even though the former have a digital humanities background and the latter a psychological one.

[1] In this work, the term *digital humanities* is mainly used as an 'umbrella term' refering to subdisciplines like digital history and computational literary studies but rather not to the many digital approaches to non-textual data.

In order to relate all three annotation layers to each other, they were all operationalized as classification tasks with both manual and automatic annotations. The latter were gained from various classifiers with a focus on "traditional" machine learning models on the one hand and transformer-based ones on the other. The actual systems were selected on the base of suggestions from the field of computational linguistics and include commonly used off-the-shelf approaches in the case of the sentiment classification. While all selected systems were compared to each other in the three classification experiments, only the approach which performed best in the respective task was considered in the case of biographical content analyses. Since this always applied to a fine-tuned transformer model which is a "black box" by nature, explainability approaches were also explored within this work as a means of making algorithmic decisions more transparent.

Content-wise, the focus of this work lies on a specific subgenre of biographical sources and a specific group of people whose lives are described in these sources. More specifically, Moravian memoirs were selected as a starting point because members of this Christian denomination have written biographical texts from the early 18th century onwards. This custom results in a comparably high amount of sources from that time which are handed down to this day. It also comprises historical accounts of elsewhere marginalized groups like peasants and women (Faull, 1992, 24). This is why Moravian sources have become the subject of inquiry not only in religious studies over the past decades. Lasch (2024, 47), for instance, stated that they "provide excellent access for interdisciplinary research to explore research topics from different scientific perspectives – linguistics, geoinformatics, cultural history, landscape architecture, botany, theology, etc. – and to understand the global impact of European expansionism". Some examples for research projects dedicated to 18th century Moravian sources are (i) "[q]uestions about [n]orms and [s]entiments" expressed in Moravian memoirs (Faull and McGuire, 2022, 125), (ii) multimodal approaches towards the spatiotemporal dimensions of the Moravian Atlantic crossings (Prell, 2022), and (iii) (post-)colonial studies of verbal devaluations (Lasch, 2023). This diversity of research topics illustrates how these sources are especially apt for research in the field of digital humanities – a field which is also characterized by its interdisciplinary orientation.

Finally, it should be mentioned that the polarity of events is prototypically modeled as an aspect level sentiment analysis in computational linguistics. This accounts for the task combination at hand. However, such an implementation still misses the chance to bridge the gap to other biographical research questions. Hence, the identification of life events and their polarities are modeled as two separate tasks here so that annotations can be analyzed on their own and in relation to each other. In addition, said task separation enables further enrichments with other annotation layers, such as the novel identification of described life stages. While the three selected classification tasks and their combinations already operationalize a broad range of potential downstream tasks of biographical research, future work could add other layers of interest which gives the proposed approach a modular nature.

1.1 Main research questions

This thesis deals with both methodological and content-related research questions where answering (some of) the former is a prerequisite for approaching the latter. The most important ones are:

Main methodological research questions

- Is it possible to overcome the domain-sensitivity of sentiment polarity?
- Is it possible to automatically infer a described life stage?
- Is it possible to automatically identify life events (grouped by the domain they concern)?
- Do transformer models outperform more transparent (rule-based) systems?
- If so, is it possible to explain their classification behavior despite their "black box" nature?
- Is it possible to overcome data scarcity issues of historical sources with supervised learning?

Main content-related research questions

- How were life stages typically depicted in 18th century biographies?
- How were life event domains typically depicted in 18th century biographies?
- Are there differences with respect to a demographic variable like subject gender?
- Were 18th century Moravian memoirs similar to other biographical sources of that time?
- Are their general patterns (all) biographical sources of that time follow?

1.2 Structure of the thesis

Chapter 2 lays the theoretical and methodological foundations of the current as well as related work by defining key concepts and explaining the most central technologies involved. More specifically, Section 2.1 starts with biographical research in general before Section 2.2 and 2.3 portray the most relevant computational approaches followed here.

Chapter 3 continues by acknowledging related work dedicated to the three text classification tasks of this thesis. To this end, Section 3.1 focuses on sentiment polarity, Section 3.2 on life stages, and Section 3.3 on life event domains. Each section starts with definitions of the underlying concepts, continues with an overview over existing computational methods and resources, and ends with issues this work tries to resolve.

The subsequent chapter 4 presents two biographical datasets the three previously described classification tasks can be applied to. The first one are Moravian memoirs (see Section 4.1) and the second one entries in a historical biographical reference work (see Section 4.2). Both datasets are compared to each other in Section 4.3 before Chapter 5 describes how these datasets were used to compile a corpus for the classification experiments described in the Chapters 6, 7, and 8.

In order to keep the experiments comparable, all three chapters follow the same structure: A first section illustrates how the aforementioned corpus was manually annotated with labels of the respective task. The second section then lists the automatic annotation approaches which are mostly the same in all three classification experiments. Hence, they are described in more detail in Section 6.2 whereas the respective Sections 7.2 and 8.2 only mention the necessary task-specific adaptions. In a third section, all automatic approaches are evaluated based on the manual annotations. To this end, the overall performance of all evaluated systems is shown first before the most promising one is evaluated further with regard to typical classification errors. As the most performant system is always a fine-tuned transformer model and hence lacking a self-explanatory decision-making process, explainability approaches are added to said error analysis. Finally, a forth section summarizes the results of each classification task.

These three classification experiments are complemented with biographical content analyses in Chapter 9 which are possible application scenarios of the classification tasks. All content analyses share their experimental setup which is outlined in Section 9.1. The first analysis looks at patterns found in Moravian memoirs (see Section 9.2) and the second one at possible gender differences within the same corpus (see Section 9.3). The third and last one adds biographical dictionary entries in order to investigate the effect of (sub)genre differences (see Section 9.4). As all three sections focus on a different content analysis, the results are always summarized in a final subsection.

Having presented all methodological and content-related results, Chapter 10 discusses the implications of these findings on a meta level. The discussion starts with the relationship between performance and resource consumption in Section 10.1, continues with more general lessons learned from the application of explainability approaches in Section 10.2, and ends with an assessment of how applicable the developed systems will be in the future in Section 10.3.

Finally, Chapter 11 concludes with an overview over the main contributions and findings of this work before pointing out possible ways for future research to build up on this thesis.

Chapter 2

Background

All fundamental concepts and methods used in later parts of this thesis are introduced in this chapter. As this thesis builds up on (digital) biographical research content-wise, Section 2.1 portrays key concepts while also summarizing typical research direction. Section 2.2 then shifts the focus to the methodology of this thesis which involves scalable reading approaches. In particular, it is also based on automatic text classifications which are the scope of Section 2.3. The subsequent Chapter 3 builds up on the overview provided in said last section by highlighting related work for each of the three classification tasks at hand.

2.1 Digital biographical research

In order to conduct any kind of biographical research – be it digital or not –, it is important to first define relevant terms, such as *(auto)biography*, *biographical dictionary entry*, and *memoir*, and to point out typical features. This is the scope of Section 2.1.1. Afterwards, Section 2.1.2 deals with computational approaches towards biographical sources – namely their digitization on the one hand and typical options for analyses on the other. In this context, it should be mentioned that no biographies were digitized within this work. Nevertheless, related concepts are still a necessary precondition for understanding the genesis of the data under investigation (see Chapter 4).

2.1.1 Biographies and similar sources

Biographies describe the life of an individual in order to highlight her or his individuality and/or social significance (Schnicke, 2022a, 5). This is why these sources are characterized by a strong subject focus and also tend to be factual in nature. However, there are certain edge cases like *fictional biographies*, such as the 1726 novel Gulliver's Travels (Gunzenhauser, 2001, 76). Similarly, biographical records are actually not required to deal with one single individual since *collective biographies* about groups of individuals exist as well (Harders and Schweiger, 2022). These deviations from a more prototypical biography will be sidelined in this thesis. The same is true for media types other than *text* which can also be biographical in nature.

Types of biographical texts

Regarding textual sources, Gunzenhauser (2001, 77) noted that "a single autobiographical text may employ formal strategies from drama, poetry, essay, and fiction". This is especially the case for 18th century sources (Werner, 2022) even though some kind of genre versatility appears to be a typical feature of biographies in general. This is even more true given that they not only appear in different media types but also in the form of diaries, journals, letters, travel reports, and many more. Most of these text types are small scale texts which Richter and Hamacher (2022) called *"biographische Kleinformen"* 'short biographies'. These prevailed in the 18th century whereas longer monographs became more common afterwards (Schnicke, 2022b; Werner, 2022). At the same time, biographical dictionaries emerged in addition to funeral and autobiographical texts. This is why the following text types are the most relevant ones for this thesis:

Dictionary entry. This text type is generally standardized within one work (Schnicke, 2022a, 6). Yet, it can also be seen as the most heterogeneous one due to conceptional differences between biographical dictionaries with some being more akin to essays and others to shorthand notes (Richter and Hamacher, 2022, 200). Due to the embeddeding in a larger work, the need for selection criteria arises. These may involve factors like importance, representativeness, and the availability of source materials (Schweiger, 2022).

Ego-document. The hyperonym of all kinds of primary sources that give insights into the life of the respective author and are hence autobiographical in nature. Examples include letters, diaries, and oral quotes (Erll, 2022, 117).

Memoir. An autobiographical account that typically has an "incremental, episodic structure" (Buss, 2001, 595). In addition, there are three further main features according to Buss (2001, 596): (i) A memoir is smaller in scale and more anecdotal than a full-fledged autobiography; (ii) "it tends to blend a number of writing practices to achieve its place as a historicized personal story, creating a style that is narrative and dramatic as well as essayistic"; (iii) the term *memoir* was used mostly synonymously with *biography* in the 18th century as it referred to "any memorial document (including obituaries)" in general at that time.

Obituary. This type of a short biography is mostly tied to funeral settings but may also appear printed in newspapers in order to inform of recent deaths (Taylor, 2001, 666). Hence, the most important defining features include (i) a recent death as a motive for writing, and (ii) a group of addressees that is public but usually related to the subject (Schnicke, 2022a, 6). Similar to this second feature, authors of obituaries tend to be rather close to the deceased one and mainly write with the goal of her or his commemoration (Richter and Hamacher, 2022, 201). Thus, they sometimes also include quotes of the subject leading to a high degree of individualization that is usually contrasted by some kind of schematization like a strict chronological record of major life events (Richter and Hamacher, 2022, 202). However, it also often results in texts which "dilute or ignore any negative aspects of the deceased's life and character" (Taylor, 2001, 667). From a typological standpoint, Schnicke (2022a, 5) stated that obituaries are difficult to classify into one of the three classical literary genres drama, poetry, and prose.

Typical features of biographical sources

No matter the biographical subgenre or text type, biographical sources can also be categorized with respect to typical features like the following:

Distance between author and subject. Sources of authors portraying her or his own life are called *autobiography* or *self-narration* whereas any greater distance is generally referred to as just being *biographic*. Therefore, the latter term may either refer to the hyperonym or antonym of *autobiographic* (Holdenried, 2022, 49). Due to the difficulty of being objective about one's own life, a greater distance tends to correlate with a higher degree of objectiveness. However, some biographers are closer to the person they write about than others which is why their accounts may involve signs of discipleship on the one hand or hostility on the other (Cockshut, 2001, 79). Concerning authors, it is also worth mentioning that "the autobiographer is living; the subject of biography is often dead" as Cockshut (2001, 78) put it. This is why it is not surprising that autobiographical accounts lack a description of the death. Interestingly, they also tend to deal less with old age while focusing on childhood and youth whereas biographies, by contrast, "pass lightly over early years" while stressing the importance of "[l]ast words" (Cockshut, 2001, 78). These different topical foci relate to the tendency that autobiographers try to illustrate how they became the person (they think) they are at the time of writing whereas biographers aim at uncovering all traits and accomplishments of their subject (Cockshut, 2001, 78).

Narrativity. This term can be defined as the depiction of a sequence of events and/or changes of state in a causal manner (Aumüller, 2022, 23–24). As such, it fits the typical depiction of life events in biographical sources which often involves some kind of retrospective sense-making (Cockshut, 2001; Holdenried, 2022). Still, these more narrative sections are typically intertwined with other elements like quotes of the subject, reflections of the author, or background information (Aumüller, 2022, 25) so that biographical accounts may be more or less hybrid.

Totality. This content-related feature assesses the amount of the life course to be covered with biographical sources usually aiming for completeness (Schnicke, 2022a, 5). Nevertheless, some are more anecdotal than others and therefore characterized by a lower degree of totality. This applies, for instance, to memoirs as noted above.

In summary, dictionary entries, memoirs, and obituaries can be compared on the base of dimensions, such as *distance* and *totality*. In this regard, the idealized Figure 2.1 illustrates that the first two are opposites of each other. Whereas a prototypical dictionary entry is characterized by a high degree of distance between the author and subject as well as a high degree of totality, a prototypical memoir is autobiographic and more anecdotal. However, due to the generic totality ambition of covering as much of the lifespan as possible, the totality difference may be less pronounced than the distance-related one. Moreover, these two dimensions are not completely independent from each other because a maximized totality requires a posthumous life narration which is impossible to achieve in the case of a true autobiographical account. This is why many biographical texts are inbetween cases like obituaries. More specifically, they are prototypically written by people closer to the protagonist than authors of dictionary entries but without the need to cover the whole life course although some do.

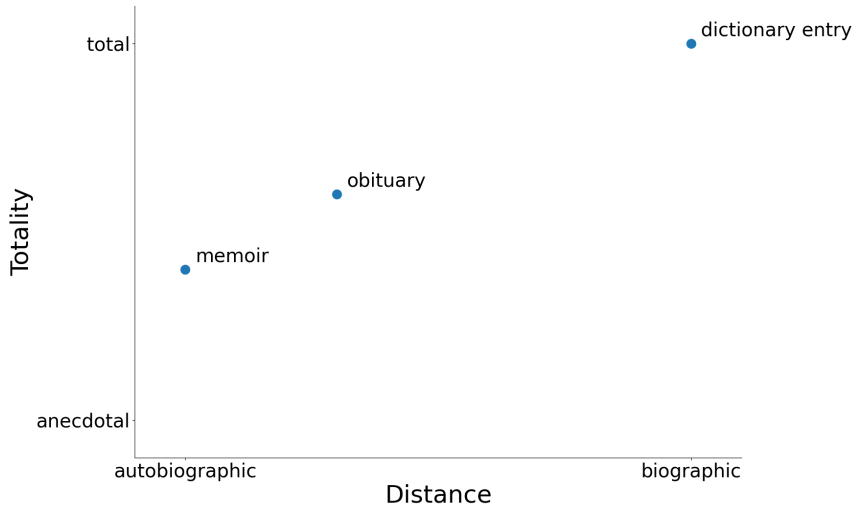

Figure 2.1: Idealized distance and totality of selected biographical subgenres.

2.1.2 Computational approaches

In digital humanities, work dedicated to biographical texts tends to focus on the digitization of raw materials and older print editions thereof. Such digitization efforts typically involve a **metadata enrichment** which may or may not follow Linked Open Data principles. **Biographical content analyses**, by contrast, are rarer in this field of research albeit that they are quite prominent in other fields, such as psychology and social studies.

Facsimiles and transcriptions

Sahle (2016, 19) claimed that source materials of scholarly inquiry are usually either (i) **"physical objects"** which are typically kept in libraries or archives or (ii) **"surrogates"** thereof. These surrogates can be further subdivided into various types the most important ones with respect to the primary sources analyzed within this thesis are the following two:

Facsimile. A surrogate covering "the visual layer by image reproduction" (Sahle, 2016, 23) "without the addition of information" (Sahle, 2016, 25). It is usually encoded in an image file format which is why it will also be referred to as a *image digitization* henceforth.

Transcription. A surrogate "on the more abstract textual layer" (Sahle, 2016, 23). In theory, this encompasses "the effort to report – insofar as typography allows – precisely what the textual

inscription of a manuscript consists of" (Vander Meulen and Tanselle, 1999, 201). In practice, however, it is up to the respective project to decide which peculiarities to encode and which not. A transcription in this editorial sense will be henceforth also called *fulltext digitization*.

These two surrogates can build up on each other in the sense of a pipeline. More specifically, it is possible to derive fulltext from image digitizations with an Optical Character Recognition (OCR) if an original textual source is typeset, or a Handwritten Text Recognition (HTR) if not. The performance of both approaches is typically influenced by (i) "the quality of the material itself", (ii) "the quality of the scanning process", and (iii) "the diversity of typographic conventions through time" (Ehrmann et al., 2023, 27:7). The last factor, in particular, may be a relevant issue in the case of historical sources and may explain lower accuracies compared to modern-day texts (Fleischhacker et al., 2025, 3–4).

Metadata enrichment and analyses

Another more abstract surrogate of any physical object is the compilation of metadata. The term *metadata* refers to "data about data" as originally first and foremost indexed by librarians (Mueller, 2014, § 24). This encompasses, for instance, structured information about the author(s), text genesis and encoding, as well as language(s) involved. In the case of ego-documents, projects typically also add demographic information about the subject of the life narration. This can include her or his name, gender as well as dates and/or places of birth and death (see Chapter 4). The reason behind this addition is that there is prototypically only one subject so that it is not unlikely that these sources may be analyzed with the intention of learning more about said individual.

In addition, some projects try to link the biographical subject to authority files so that further metadata fields can be added by reconciling (Blessing et al., 2015; Schlögl and Lejtovicz, 2017; Hyvönen et al., 2022). However, this is difficult if not impossible in the case of historical sources about less well-known people because they may not have an entry in authority files.

Still, metadata fields can also be reused in a faceted search that simplifies the selection of relevant subcorpora for further inquiry (Reinert et al., 2015, 15). Moreover, especially demographic metadata fields sometimes even become a subject of inquiry on their own (Bernád and Kaiser, 2018; Leskinen et al., 2018).

Biographical content analyses

Most biographical content analyses conducted within the field of digital humanities focus on (automatically) annotated entities, such as persons and places, and typically involve *network analyses* (Stotz et al., 2015; Bernád and Kaiser, 2018; Faull, 2021). More topic and/or event-related studies tend to deal with modern-day biographies found on Wikipedia due to the scarcity of other biographical sources available off-the-shelf in digital form (Bamman and Smith, 2014; Palmero Aprosio and Tonelli, 2015; Stranisci et al., 2023). Both these lines of research will be henceforth sidelined because they differ significantly from the content analyses conducted within this thesis due to (i) a more recent time of writing and (ii) the selected methodological approach.

However, there are also a few studies like the ones of McGuire (2021) and Dennis-Henderson et al. (2020) which are more relevant for this thesis. Both studies try to assess the depiction of topics described in the respective historical biographical sources under investigation. To this end, they both apply sentiment analysis techniques (see Section 3.1.3) although they take different methodological approaches for the identification of the sentiment targets. More specifically, McGuire (2021, 130) identified topic-related passages with keyword searches whereas Dennis-Henderson et al. (2020, 93) relied on a topic model.

As hinted at above, the situation is different in psychology and social studies where a lot of work exists that relies on the computational and/or computer-assisted analysis of autobiographical interviews no matter if the analysis is conducted quantitatively (Haehner et al., 2024) or qualitatively (Schubring et al., 2019). In the context of sociological interviews, subjects are sometimes asked to draw so called **life satisfaction charts** which follow the (perceived) ups and downs of their life chronologically and are labeled with major life events (Clausen, 1998, 200–201). This approach

shares some similarities with the sentiment trajectories analyzed in Chapter 9. Nevertheless, it differs from the one followed in this thesis with regard to (i) the underlying concept, (ii) the process leading to the respective chart, and (iii) the different media type of the source material. Moreover, it is worth mentioning that autobiographical interviews are never total – i.e. cover the whole lifespan of an individual from birth to death – because the subject has to still be alive to be interviewed. Biographical records in written form, by contrast, aim at totality whenever possible – i.e. the death of the biographical subject occurred before the text finalization – and the ones analyzed in later chapters are all expected to achieve it. Still, at least in theory, both types of sources can be analyzed with respect to different scales in the sense described in the subsequent section.

2.2 Scalable reading

> "Digital tools and methods certainly let you zoom out, but they also let you zoom in, and their most distinctive power resides precisely in the ease with which you can change your perspective from a bird's eye view to close-up analysis. Often it is a detail seen from afar that motivates a closer look."
>
> Mueller (2014, § 31)

This quote summarizes the advantages of a mixed methods approach often called *scalable reading* which the methodology employed in this work follows on the most abstract level. Said approach combines a qualitative close inspection of noteworthy singularities (= **close reading**) with a quantitative assessment of global trends on a corpus level (= **distant reading**). While these terms are in widespread usage, Jockers (2013, 25) preferred the terms *micro-* and *macroanalysis* since especially the latter "is no longer reading" in its prototypical meaning.

Close reading. Jänicke et al. (2015, 84) defined this first approach as "the thorough interpretation of a text passage by the determination of central themes and the analysis of their development." Besides the actual act of reading, it typically involves highlighting and commenting textual units deemed important or otherwise noteworthy. However, this practice is only applicable to a limited amount of textual units in practice since "time is still a limiting factor for close reading" (Elwert and Gerhards, 2017). Hence, it is also "a largely qualitative approach" where inferences are based on a few selected cases (Jockers, 2013, 25).

Distant reading. Underwood (2017, § 5) defined this second approach as "the practice of framing historical inquiry as an experiment, using hypotheses and samples (of texts or other social evidence) that are defined before the writer settles on a conclusion". In this sense, a methology from the social sciences is applied to (computational) literary studies. Moreover, the hypothesis-driven experimental setup aims at "the reduction of researcher bias" (Dennis-Henderson et al., 2020, 90). Typical sample sizes "range from a few dozen instances to a million or more" and need to be "constructed" (Underwood, 2017, § 10). This (sub)sampling or *corpus construction* enables the identification of "overall patterns" (Dennis-Henderson et al., 2020, 90) which are usually not inferable from individual texts alone. In theory, it even allows for sample sizes close to a full survey and thereby circumvents canonicity issues (Moretti, 2005). In practice, current applications include an increasing amount of methods from corpus and computational linguistics as well as visualization techniques so that they have become even more interdisciplinary in nature. The underlying reasoning relates to the observation that one can "learn a lot about a large corpus if each of its analyzable items is consistently described in terms of very few and quite simple data points" as Mueller (2014, § 28) put it. These data points can be metadata fields but also (corpus) linguistic units. No matter the exact implementation, the result is always an "abstract view" that may involve means of data visualization (Jänicke et al., 2015, 84–85).

Scalable reading. Mueller (2014, § 32) saw the "true power" of this third approach in the possibility to uncover "the shared verbal patterns that make up a genre or tradition." While this is also true for distant reading alone, a fuller picture may emerge from a workflow that uses the latter for "initial and subsequent explorations" in order to identify outliers suitable for close reading

(Coles and Lein, 2013, 152). In this way, the advantages of one approach may circumvent the drawbacks of the other. In addition, it accounts for the fact that "direct access to the source texts is important for humanities scholars" (Jänicke et al., 2015, 85).

Besides these scales, Elkins (2022, 46) introduced a "clustered close reading" of the (visual) outputs of multiple different systems applied to the same dataset in order to identify notable text sections. This approach is especially apt for tasks that can be operationalized not only with manual annotations but also with the many different automatic approaches outlined in Section 2.3.

Corpus linguistic units

Distant reading approaches in general and text classifications in particular are based on various basic (corpus) linguistic units. The most important ones in the context of this work will be used in the following senses henceforth: A **corpus** is seen as the largest textual unit under consideration and consists of **documents** which are typically a single *text* like a biography. **Tokens**, by contrast, are the smallest unit under consideration and typically a *word* in the linguistic sense. While *tokens* account for any word form, **types** refer to unique ones. Similarly, a **lemma** is a further abstraction as it refers to the "basic" form of a word as listed in a dictionary. Finally, the term **n-gram** is used for a sequence of units, such as words or individual characters, that share a fixed length `n`.

Term frequency (TF) and inverse document frequency (IDF)

One common way to assess the aforementioned units in an information retrieval sense is to count the (relative) number of occurrences (= **TF**). As some terms are quite common in general and hence less relevant for most downstream tasks, the TF is commonly multiplied by an **IDF**. Given a collection `D`, a term `t`, and a document $\mathtt{d_t}$ comprising said term, this measure can be formalized as follows (Leskovec et al., 2019, 9–10):

$$IDF_t = log_2(\frac{D}{d_t}) \tag{2.1}$$

2.3 Automatic text classification

On a more practical level, the methodological approach of this thesis is based on what is known as **text classification** – i.e. the assignment of categorial or discrete variables in a statistical sense to textual units. In this regard, it shares similarities with the aforementioned close reading annotations but is more restricted: For instance, the classes to be annotated are limited in number and defined a priori. Moreover, the textual units are also predefined and may only be annotated with one single label at a time. These latter two constraints distinguish a text classification from a **token classification** where not the whole textual sequence but each individual token is attributed a label (or the lack thereof). Moreover, the single label condition excludes a multi-label classification which is, in principle, also possible in a close reading setting as well as in the case of automatic annotation algorithms but henceforth sidelined.

Target-wise, automatic text annotations serve common use cases as identified by Bamman et al. (2024, 496–498) in this work. These are first and foremost (i) an evaluation of **category coherence**, (ii) a **category sensemaking**, (iii) a **search for outliers**, and (iv) an **upscaling** from manual annotations to larger samples. While the first three play a more implicit role in various sections to follow, the last one is the base of the content analyses presented in Chapter 9.

One benefit of an upscaling is the temporal acceleration of the annotation procedure although it comes at the cost of a lower annotation quality and potential bias issues. This is why different automatic annotation approaches are compared to each other in the current work and a more technical portrayal of the differences between them is the scope of Section 2.3.1. As all selected systems follow the single label condition, ways to handle ambiguous cases are necessary and described in Section 2.3.2. Afterwards, Section 2.3.3 lists all evaluation metrics used within this

work. Finally, Section 2.3.4 concludes by showing how especially the decision-making processes of deep learning approaches can be approximated.

2.3.1 Automatic annotation approaches

There are different ways to automatically annotate biographical events or sentiment polarities in particular, or operationalize text classification tasks in general. System-wise, there is a continuum spanning from **dictionary lookups** over **"traditional" (supervised) machine learning** approaches to **transformer-based language models**. Different systems from this continuum which are relevant for this thesis are portrayed in the following subsections in said order.

Dictionary lookups

In an automatic text classification setting, the term *dictionary* refers to a list of key value pairs. The **keys** are usually tokens or lemma forms while the **values** can be class labels, single numbers, or n-dimensional vectors. If the task at hand requires a different output, dictionary values need to be postprocessed which is often accomplished by some kind of mapping. The most straightforward approach to resolve this issue in a setting where the raw output values are continuous and the target classes discrete is rounding.

Since dictionaries may vary with regard to the amount of keys they contain, the applicability of a given dictionary to a dataset which needs to be annotated can be assessed to a certain degree a priori by measuring the expected *coverage* (Herrmann and Grisot, 2022). In practice, this is usually operationalized by calculating the relative number of word forms `W` in the dataset which are also part of the dictionary keys `K`. This can be formalized with Equation 2.2 and the weighting may be tweaked – for instance by preprocessing word forms.

$$Coverage = \frac{|W \cap K|}{|W|} \tag{2.2}$$

It should be noted that while this assessment can be used to compare different dictionaries, it is not an actual evaluation metric like the ones listed in Section 2.3.3. This is due to the fact that it is impossible to compare coverages between tasks and datasets since a low amount of task-related vocabulary may just be a feature of a given dataset.

"Traditional" machine learning approaches

On a macro level, machine learning approaches can be roughly divided into supervised and unsupervised methods depending on whether a labeled dataset is available for training. Such a dataset is usually either compiled manually or automatically derived from available resources. In a *supervised* setting, it is then used to train a machine learning model, whereas *unsupervised* approaches are not trained to the task and/or data at hand. Some examples include Decision Trees, Naive Bayes, or Support Vector Machines (SVMs). However, all these system types have received significantly less attention in related work after the emergence of deep learning methods in general and the transformer architecture in particular which is why they are called "traditional" here. Furthermore, their usage shifted, especially in the field of computational linguistics, to some kind of baseline status more promising systems should outperform in a system comparison. Especially in the case of a sentiment classification, which is the only established classification task of this work, SVMs are the most widespread ones in this manner (see Section 3.1.2). This is why they were selected in order to represent "traditional" machine learning approaches in general. The underlying architecture and the basic functioning of any SVM are portrayed in the following.

SVMs

SVMs are non-probabilistic machine learning systems originally proposed by Cortes and Vapnik (1995) for binary classification tasks. However, current implementations enable an application to multi-class settings, too. The approach is based on *support vectors* which are the data points in a

feature space lying closest to a decision border or *hyperplane* (see Figure 2.2). In simple terms, an SVM does not consider all training samples but only the ones deemed most discriminatory. This is part of the reason why their training is rather computationally inexpensive.[2] In order to build a feature space, some kind of feature selection or vectorization is necessary. In addition, there are different kernel functions which are not explored in detail here. Instead, only a linear kernel was used within this work because this setting was found to be promising in related work as outlined in Section 3.1.2.

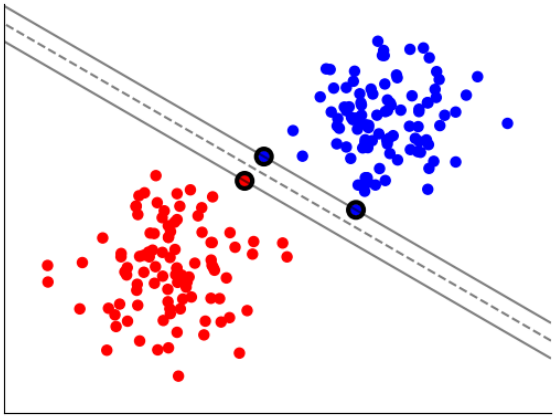

Figure 2.2: Example of a binary SVM classifier (samples of the first class in red and the second in blue, the support vectors are enclosed in circles, the dotted line represents the optimal hyperplane and the solid ones the optimal margin; based on Cortes and Vapnik 1995, 275)

Embedding text into numerical vectors

As machine learning models in general and SVMs in particular typically require numerical vectors as input, the textual input of a text classifier needs to be vectorized first. This is called *feature selection*, and when relying on distributional features *text embedding*. Common distributional approaches include (i) a binary presence vectorization, (ii) a count vectorization, and (iii) a TF-IDF vectorization. This is why they are included in the widespread Python module *sklearn*[3]. In all three cases, a **vocabulary** is built first and specifies the number of features or dimensionality of the vector space. This vocabulary is typically either based on character or word n-grams. As implied by the names, a binary presence vectorizer then checks for each term of the vocabulary if it occurs in a given instance or not, a count vectorizer weighs them with their TF, and a TF-IDF vectorizer with TF-IDF values. In summary, this means that the main difference between these three common distributional vectorization approaches is the respective weighting involved. The vocabulary and therefore dimensionality, by contrast, depends mainly on the training dataset as well as preprocessing steps. Hence, it is possible that the application to unseen data during inference leads to scarce dimensions if certain n-gram features are rare, or even out-of-vocabulary issues if the training data lacks potentially important features.

To at least mitigate the problem of an unbound vocabulary which tends to lead to scarce dimensions, more recent approaches rely on a fixed set of **subword tokens** instead of actual words. This design choice may also facilitate the learning of word formation patterns. In addition, the embeddings are "contextualized" (i.e. depend on surrounding tokens) and often include dimension reduction steps. However, especially the latter two extension levels which are a common backbone of deep learning settings come at the cost of a higher resource consumption but tend to be less useful for simpler machine learning models like SVMs (Laurer et al., 2024).

[2] Laurer et al. (2024, 96) stated that it usually takes only a few "minutes on a Laptop CPU".

[3] https://scikit-learn.org/stable/modules/feature_extraction.html (Last accessed on February 18, 2026)

Deep learning approaches

Different from more "traditional" machine learning approaches, Deep learning systems involve some kind of *artificial neural network* with input, hidden, and output layers. Typical deep learning architectures are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory systems (LSTMs). Although they are still in use, **transformer** models (Vaswani et al., 2017) have become the backbone of most, if not all, current natural language processing systems. The transformer architecture is illustrated in Figure 2.3 and shows that it involves text embeddings and a stack of hidden layers. The input embeddings are contextualized since the position of an input token in the input sequence is also encoded. Each hidden layer is composed of a Multi-Head Attention mechanism and a fully-connected Feed Forward network followed by a normalization. The output layer is a linear one with a softmax function and generates the probabilities for predicting the output classes of the respective task at hand.

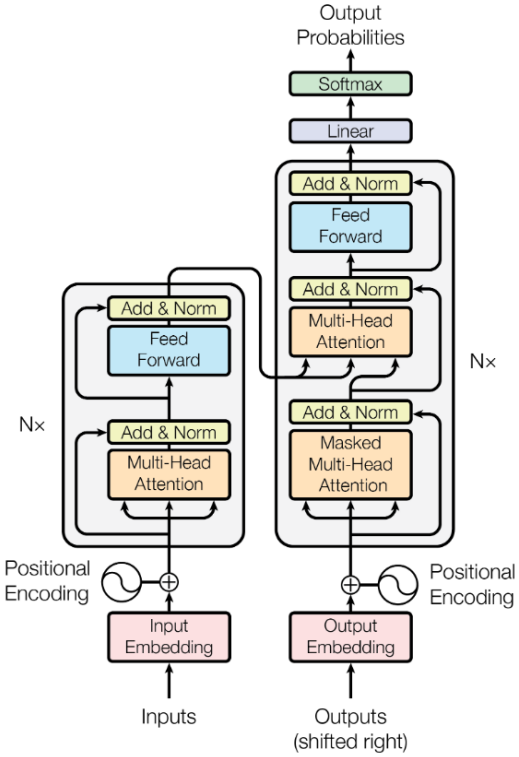

Figure 2.3: The transformer architecture (Vaswani et al., 2017, 6001)

The current ubiquity of transformer models is underlined by 1 million models for approximately 50 generic tasks and 5000 language varieties on the main dissemination platform called *Hugging Face Model Hub* (Wolf et al., 2020). One of the most widespread families of transformer-based language models are Bidirectional Encoder Representations from Transformers (BERT):

BERT models

BERT was proposed by Devlin et al. (2019) and the respective architecture is composed of bidirectional transformer encoder blocks as implied by the name. BERT models are **pre-trained** with vast amounts of data in order to gain "semantic" *world knowledge* that can later be exploited for a task-specific fine-tuning. The pre-training is conducted in a multi-task manner as BERT models are simultaneously optimized for solving the following two tasks:

Masked Language Model. In this task, 12 % of the input tokens are randomly selected to be replaced with the special token `[MASK]` and another 1.5 % with a random one. The model then needs to predict the original unmasked token.

Next Sentence Prediction. This binary classification task uses the two labels `IsNext` and `notNext` as well as a sentence pair input. One half of the training data are sequences of originally

sequential sentences whereas the other half are single sentences followed by a random but unrelated one from the training data corpus.

This pre-training is called *self-supervised* as the training data can be generated from basically any dataset. It has also been shown to lead to emergent features like the possibility to be fine-tuned to most if not any task(s) at hand. This will be shown in later chapters. Additionally, BERT models seem to gain an implicit understanding of linguistic structures (Manning et al., 2020).

Fine-tuning is a special transfer learning setting where task-specific input and output layers replace the original ones and all parameters are updated end-to-end in the case of BERT models (Devlin et al., 2019, 4175). Although a full fine-tuning requires less amounts of computational resources than the pre-training, the resource consumption is still notable. Hence, various alternatives have been proposed over the past years and most of them are subsumable under the term *Parameter-Efficient Fine-Tuning (PEFT)* (Ding et al., 2023).

Natural Language Inference (NLI) models for zero-shot classification

A zero-shot classification "consists in recognising new categories of instances without training examples" according to Romera-Paredes and Torr (2015, 2152). It is thus some kind of unsupervised transfer learning setting where a classifier sees all possible class labels at inference time only. Besides architectural details, zero-shot systems tend to differ with regard to the question on how to present the class choices with a simple label list on the one side of a continuum and (un)structured class descriptions on the other.

One way to accomplish this in practice is by training a machine/deep learning model on the NLI task which is sometimes called *entailment approach* (Yin et al., 2019). NLI is a ternary classification task with the labels `contradiction`, `neutral`, and `entailment`. Different from more prototypical text classification tasks, the input is not a single text but a `context – hypothesis` pair. In a zero-shot setting, this means that each input sequence to be annotated becomes the *context*, whereas the labels of the task at hand form the *hypotheses*. Then, the input is subsequently tested against each of the hypotheses (i.e. downstream task labels) and the best fit is chosen as the final output.

Large Language Models (LLMs)

Recently, generative LLMs have emerged as a second promising zero-shot classification approach besides NLI models. This is why Zhang et al. (2025), for instance, grouped zero-shot models in general based on the amount of model parameters into *prompt-based* LLMs on the one hand and smaller transformer models fine-tuned for NLI on the other. While NLI models are typically based on class label lists, LLM classifiers tend to operate on descriptions in natural language instead. This peculiarity explains why LLMs are attributed a higher degree of "accessibility" than other deep learning approaches (Klähn et al., 2025, 2).

It is worth mentioning that performances in zero-shot settings seem to correlate with the number of model parameters to some extent. These abilities are hence seen as *emergent* features of LLMs (Brown et al., 2020). This may be part of the reason why the number of model parameters has increased over the past years so that the threshold of what counts as "large" is somewhat fluid.

2.3.2 Ambiguity handling

Strategies for dealing with ambiguities are a core aspect of (reflected) text analyses as many real-world datasets comprise instances which may be related to multiple classes. Such a handling can be either implemented a priori with respective remarks in the annotation guidelines or a posteriori by postprocessing the raw classifier output.

One way to deal with ambiguities **a priori** is to allow for the assignment of multiple labels and to delegate the reduction to a single label to some kind of postprocessing (see e.g. Mostafazadeh Davani et al., 2022). As an alternative, *tie-breakers* can be defined in the annotation guidelines in order to assure that only a single label is attributed per instance. Another possibility is the introduction of

residual classes, such as `mixed` (Davoodi and Mezei, 2022, 219) for cases which are ambiguous due to being relatable to multiple classes, on the one hand, or `not clear` (Haehner et al., 2024, 9) for truly ambiguous ones, on the other.

An **a posteriori** handling of ambiguities may involve an aggregation strategy, such as a *majority vote* in the case of multiple annotations from different annotators (Mostafazadeh Davani et al., 2022). In the case of classifiers based on machine learning, a softmax function over the per-class probabilities is commonly used in a similar manner (see Figure 2.3).

It is also possible to factor in the *output probabilities* or *confidence scores* as a weighted mean if the categorial class labels can be converted to continuous values. Rebora et al. (2023, 40), for instance, followed this approach in the case of a sentiment classification. Furthermore, the per-class confidence scores can also be exploited to model cases an automatic classifier "considers" ambiguous because there is a tie or close tie between multiple class labels. In such a case, a disambiguation strategy may again be necessary for downstream tasks or system evaluations. In the latter case, it even remains an open research question "how to deal with [...] 'real' ambiguities" according to Pagel et al. (2018, 35).

2.3.3 Evaluation

There are various ways to evaluate automatic annotations (Opitz, 2024) and most of them are based on the comparison with some kind of (manually compiled) ground truth. In the simplest form, which is a binary classification, this results in the four possible cases depicted in Table 2.1.

True label	**Predicted label**	
	1	0
1	TP	FN
0	FP	TN

Table 2.1: True/false positives (TP/FP) and negatives (TN/FN) in a binary classification setting

A "correct" classification means that an automatically *predicted label* equals the *true label* found in the ground truth. This applies to **true positives (TP)** as well as **true negatives (TN)**. The former can be related to all items of a ground truth `y` in order to calculate the proportion of "correct" classifications which is called **accuracy** (see Equation 2.3) (Opitz, 2024, 823).

$$Accuracy = \frac{|TP|}{|y|} \tag{2.3}$$

Especially in cross-task comparisons, this score can be related to the aforementioned *category coherence* use case whenever one "employs classification accuracy as a proxy for the coherence of a concept itself, under the assumption that statistical regularities could only exist for concepts that are somehow real" (Bamman et al., 2024, 497). This means that a lower accuracy is to be expected in the case of one or more less coherent categories and implies that classification experiments may hint at disparities of this kind.

Other evaluation metrics, such as **precision** (see Equation 2.4) and **recall** (see Equation 2.5), focus on either **false positives (FP)** – i.e. instances which are falsely labeled with the target class – or **false negatives (FN)** – i.e. instances which are not found (Opitz, 2024, 821). In this context, it is important to stress that it depends on the downstream task at hand which one of these errors is more problematic than the other.

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{2.4}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{2.5}$$

In order to account for cases where both errors should be considered (almost) equally important, the harmonic mean of precision and recall which is called **F1 score** can be calculated with Equation 2.6 (Opitz, 2024, 825). Please note that a harmonic mean (i.e. F1 score) equals its components (i.e. precision and recall) by definition if these have the same value.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times |TP|}{2 \times |TP| + |FP| + |FN|} \tag{2.6}$$

When applying precision, recall, or the F1 score to a multi-class classification setting, the scores are calculated for each of the class labels in the set `L` separately in a one-vs-rest fashion and subsequently averaged to an overall score. The most common averaging strategies are called **micro**, **macro**, and **weighted**. The first one is calculated with the sums of the per-class TPs, FNs, and FPs (see Equation 2.7). It should be noted that FNs of one class are FPs of the others and vice versa which is why their sums average out meaning that a micro precision, micro recall, and by extension micro F1 score are always identical. Moreover, the sum of per-class TPs and FNs equals the number of instances with the respective true label $\mathtt{y}_l$ by definition (see Equation 2.8) which is why all micro averaged scores are also always the same as the accuracy (Opitz, 2024, 823):

$$Score_{micro} = \frac{\sum_{l \in L} |TP_l|}{\sum_{l \in L} |TP_l| + \sum_{l \in L} |FP_l|} = \frac{\sum_{l \in L} |TP_l|}{\sum_{l \in L} |TP_l| + \sum_{l \in L} |FN_l|} = \frac{\sum_{l \in L} |TP_l|}{|y|} \tag{2.7}$$

$$|y_l| = |TP_l| + |FN_l| \tag{2.8}$$

The second averaging strategy (= **macro**) is the arithmetic mean of the per-class scores (see Equation 2.9). It may be favorable in cases where all classes (and especially the rarer ones) are equally important since they are equally assessed by the score (Opitz, 2024, 824–825):

$$Score_{macro} = \frac{\sum_{l \in L} Score(l)}{|L|} \tag{2.9}$$

A **weighted** average is calculated by factoring in the respective proportion of instances in the ground truth (see Equation 2.10) thus accounting for class imbalances in the ground truth dataset (Opitz, 2024, 825). It is noteworthy that a weighted F1 score does not have to be a value between the weighted precision and recall. Furthermore, a weighted recall is always the same as the micro averaged scores and accuracy due to 2.8.

$$Score_{weighted} = \frac{\sum_{l \in L} |y_l| \times Score(l)}{|y|} \tag{2.10}$$

$$Recall_{weighted} = \frac{\sum_{l \in L} |y_l| \times \frac{|TP_l|}{|TP_l| + |FN_l|}}{|y|} = \frac{\sum_{l \in L} |TP_l|}{|y|} \tag{2.11}$$

Last not least, it should be noted that there are further ways to evaluate automatic text classification systems. Pichler and Reiter (2022, 12–16) offered a categorization which explicitly targets digital humanities research. They call the aforementioned performance metrics **reliability** measurements following the terminology of Krippendorff (2019) who deals with different types of **validity**, as well.

While performance metrics are quantitative in nature, other performance assessments are more on the qualitative side. This applies, for instance, to typical error analyses where selected misclassifications are investigated in a close reading setting (see Sections 6.3.2, 7.3.2, and 8.3.2). In addition, there are ways to implement an **interpretability** assessment which are explored further in the following section. They are especially important in cases where datasets "are not representative or small, or performance metrics low" according to Pichler and Reiter (2022, 15). In a similar vein, Bamman et al. (2024, 497) suggest to scale annotations only up "[a]fter establishing [both] the validity and interpretability" of a selected classification system.

2.3.4 Explainability, interpretability, and similar concepts

While rule-based automatic annotation approaches like dictionary lookups are inherently explainable and interpretable, ones based on machine learning and even more so deep learning are not. This is why the latter are also called "black box systems" which is a major problem in fields like healthcare or automated means of transport where every decision-making matters (Linardatos et al., 2020, 2). Hence, it is not surprising that ways to overcome this issue have been under investigation since the early days of artificial intelligence (AI) research in the mid-1970s (Adadi and Berrada, 2018, 52139). In the field of digital humanities, this issue has recently gained some importance, too, due to the rise of deep learning approaches where system transparency and trust issues can be raised.

Whereas *transparency* is some kind of umbrella term, *explainability* and *interpretability* tend to be more clearly distinguished from each other with Definition (2.1) followed here in the case of explainability. Please note that interpretability approaches will be henceforth sidelined as they aim at a more technical understanding of "the internal structure and functioning of the model" under investigation (Şahin et al., 2025, 868).

(2.1) *Explainability refers to a model's ability to provide clear and understandable information about its predictions and decisions.*
(Şahin et al., 2025, 868)

In their overview over existing explainability approaches, Linardatos et al. (2020, 5) grouped them by different dimensions. One dimension is based on the question of whether explanations are **local** (i.e. explain the classification of a single instance), or **global** (i.e. explain the general decision-making with regard to a whole dataset). Another more technical grouping dimension is that some approaches are **model-specific** while others are **model-agnostic**. In the case of transformer-based classifiers, one model-specific approach is an attention visualization. Model-agnostic approaches, by contrast, are applicable to most if not all machine and/or deep learning models. Some of the currently most promising approaches are a Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). The first and last one are used within this work and hence portrayed in the following.

LRP

LRP is a model-specific explainability approach which involves two phases: First, a neural classifier performs a regular forward pass from the input through hidden layers to generate the output (probabilities). Then the contribution of individual features to the final prediction are traced backwards through the neural network in a layer-by-layer fashion – as implied by the name *Layer-wise Relevance Propagation*. This is usually implemented with some kind of Deep Taylor Decomposition (Montavon et al., 2017).

This explainability approach was originally proposed for image classifiers by Bach et al. (2015) but soon also adapted to text classification tasks like sentiment analysis (Arras et al., 2017). System-wise, Bach et al. (2015) applied LRP to SVMs and CNNs, and Arras et al. (2017) to RNNs. In addition, Chefer et al. (2021) offered with *Transformer-Explainability* an LRP-based approach for transformer models.

In practice, all these model-specific variants have in common that LRP assigns **relevance scores** which are usually normalized to the range `[-1,+1]` with zero representing "irrelevant" features. Negative values stand for features which oppose the output class considered. However, some implementations, such as Transformer-Explainability, ignore negative values (Chefer et al., 2021, 3–4) in order to account for peculiarities of the system type to be explained.

These scores are then commonly visualized with heatmaps in order to highlight the relevance of input features – be they pixels or textual tokens – on the final model output. Moreover, relevance scores can be used to investigate the more general model behavior by analyzing multiple input-output pairs in order to approximate a global explainability. Especially this latter practice can be related to the aforementioned *category sensemaking* use case of automatic annotations which

Bamman et al. (2024, 496) define as "understanding the characteristics of a category through the features that are predictive of it" and will be followed in Sections 6.3.3, 7.3.3, and 8.3.3.

SHAP

Lundberg and Lee (2017, 2–3) developed SHAP for measuring feature importance in an additive manner by combinining the six prior methods (i) LIME, (ii) DeepLIFT, (iii) LRP, (iv) Shapley regression values, (v) Shapley sampling values, and (vi) quantitative input influence. As the last three components and hence half of them involve the computation of Shapley values from cooperative game theory, this local explainability approach is called *SHAP* and the respective Python module `shap`. The additive nature is shown in Figure 2.4.

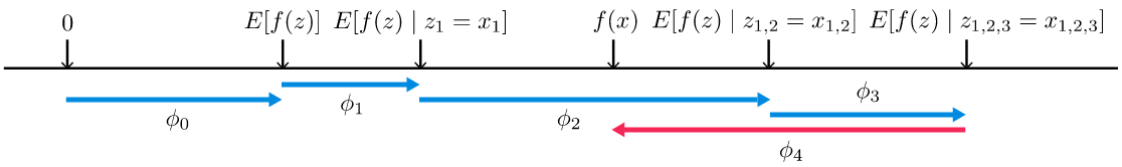

Figure 2.4: Basic SHAP explanation plot for a classifier with some bias towards predicting the target class z (= ϕ_0) when applied to an input with three features ($\phi_{1\text{-}3}$) supporting the final prediction and one (= ϕ_4) opposing it (Lundberg and Lee, 2017, 5)

As illustrated by Figure 2.4, SHAP values measure the impact `E` of a feature ϕ – for instance, a subword token in the case of BERT models – on the final prediction probability `f(x)`. These values can be either positive if they support the final prediction like in the case of the abstract features ϕ_{1-3} or negative if they oppose it (see the direction of the arrow for feature ϕ_4). ϕ_0 is different from the actual features as it accounts for the fact that statistic classifiers tend to favor a certain class label when lacking input features. This kind of *bias* is called *base value* `E[f(z)]` in the case of SHAP. It is increased by positively impacting input features and decreased by negatively impacting ones until the final prediction probability `f(x)` is reached.

Different from LRP relevance scores, **SHAP values** are normalized in three ways: First of all, they range from -1 to +1 with the algebraic sign standing for the polarity of the impact. Secondly, the sum of all SHAP values per feature equals the probability that the model under investigation predicts a given class label for a given input sequence. Finally, the sum of all per-class SHAP values is always 0 for any given input feature because a feature supporting the prediction of one or more classes has to also oppose the prediction of others to be distinctive.

All these properties enable a potential upscaling from single instances to a whole dataset in order to reach a global instead of a local explainability. However, it should be added that the "exact computation of SHAP values is challenging" according to Lundberg and Lee (2017, 5). This is why they and their Python implementation approximates actual values although said approximation is still quite resource-intensive (Zielinski et al., 2023; Brookshire and Reiter, 2024).

Chapter 3

Related work

The main focus of this work is to conduct three classification tasks on historical biographical sources in order to identify notable patterns. Two tasks have not yet been operationalized in the exact same way as outlined in the following chapters. Still, there is a lot of work this one can build up on. First and foremost, this applies to content analyses based on a sentiment classification as they have become widespread in many disciplines. This is why this chapter starts with sentiment analysis and a focus on the historical domain (see Section 3.1). Section 3.2 follows with possible subcategorizations of the life course before Section 3.3 concludes with the extraction and categorization of biographical events. This order also follows the continuum from a more general task to more biography-specific ones. All three sections to follow are structured in a similar way: They start with definitions of the task and related concepts in a first subsection and continue with an overview over computational approaches in a second one. Due to the ubiquity of sentiment analyses, a penultimate subsection deals with actual downstream tasks in digital humanities which has no counterpart in the other two sections. Despite this difference, all three sections end with a summary of typical task-specific issues as well as methodological critique raised in related work.

3.1 Historical sentiment analysis

Sentiment analyses have traditionally been part of the canon of methods from social sciences but have also been embedded in interdisciplinary research for some time now. This is especially true for the field of digital humanities, as inferable from the fact that a simple search for "sentiment" yielded results in 234 research articles published in the journal *Digital Scholarship in the Humanities* and 100 in *Digital Humanities Quarterly* in early February 2026. In addition, this kind of analysis has "become one of the most discussed topics" in the partly overlapping fields of computational literary studies (Rebora, 2023) and "one of the most widely studied tasks" in natural language processing (Roccabruna et al., 2022, 581). Still, even more related work falls into the field of computational linguistics as more than 47,000 results in the ACL anthology underline. In light of these numbers, literally any survey on related work needs to disregard a lot of research endeavors even if, for instance, focusing on the historical domain. Nevertheless, the following first Subsection 3.1.1 highlights the most important steps from the early days of computational polarity analysis to the current variety of methods subsumed under the term *sentiment analysis*. This conceptual variability leads to a variety of methods presented in Section 3.1.2 as well as typical downstream tasks (see Section 3.1.3). Section 3.1.4 lists the most important issues that need to be resolved when conducting a sentiment analysis in general and applying it to the historical domain in particular. This final section also summarizes methodological critique in order to stress that especially the high degree of variability of underlying concepts and implementation details requires various kinds of reflections to increase the validity of findings gained from a computational sentiment analysis.

3.1.1 Polarity, sentiment, and similar concepts

Sentiment analysis as a computational classification task can be traced back to the prediction of the **semantic orientation** or **polarity** of adjectives (Hatzivassiloglou and McKeown, 1997). Said predecessor approach relied on a seed list of adjectives known to be rather positive (e.g. *reputed*) or negative (e.g. *troublesome*) which should be extended based on conjunctions in a corpus. The original downstream task in mind was the identification of semantic relations like antonymy and synonymy. Yet, it soon changed after Hatzivassiloglou and Wiebe (2000) observed that polar adjectives may be promising features for a sentence-wise **subjectivity** classification which differentiates between subjective phrases, on the one hand, and factual ones, on the other.

One of the earliest studies fully dedicated to **sentiment** on the level of whole texts was conducted by Das and Chen (2007) and dealt with stock board messages. Their methodology also involved the identification of adjectives, followed by a polarity dictionary lookup and steps to aggregate these values from the word over the phrase to the message level (Das and Chen, 2007, 1379). While aggregating, they added a `neutral` class. Furthermore, they grouped the annotated messages by day in order to align their **sentiment series** with the time series of a stock index (Das and Chen, 2007, 1384). At that time, sentiment analyses were in fact not limited to the financial domain. Pang et al. (2002, 79), for instance, identified an added value in categorizing on-line articles by sentiment due to the "rapid growth in on-line discussion groups and review sites".

Definitions of sentiment analysis

One of the first definitions of the term *sentiment* is the following:

(3.1) *sentiment, or overall opinion towards the subject matter*
(Pang et al., 2002, 79)

Whereas the two terms *sentiment* and *polarity* were initially used somewhat interchangeably, the former subsequently gained more senses. On the one hand, it has become common to refer to quite different tasks with the umbrella term *sentiment analysis* which has been subject to methodological critique (see Section 3.1.4). On the other hand, it led to broader definitions that try to be as inclusive as possible. One example for the latter is the following:

(3.2) *Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes.*
(Liu and Zhang, 2012, 415)

The list in the end of this second definition can be seen as just some kind of subcategorization of the "subject matter" in the aforementioned one. The first part hints at the existence of task name variants, such as "opinion mining". In a similar vein, the middle part (= "study of people's opinions, appraisals, attitudes, and emotions") shows the high degree of variability with regard to the underlying concepts. Venkit et al. (2023, 13744) called them sentiment "frameworks". Since *opinions* and *emotions* can be seen as the most common ones, Yadollahi et al. (2017, 25.2) used them as the main subcategories in their taxonomy. It is depicted in Figure 3.1 and illustrates that *polarity* is in fact independent from this subcategorization. More specifically, both *opinions* and *emotion categories* can be bipolar. One example for the latter is the *Valence-Arousal* model which is actually multi-dimensional but still based on two polarities (Kim and Klinger, 2021). This line of research as well as other subtasks of opinion mining will be henceforth sidelined.

Sentiment polarity classification

The rather open definition also leads to a wide range of different implementation details. Still, the main variants of a *polarity* classification have in common that they deal with quantitative variables which are either discrete or continuous in statistical terms. The simplest form is a **binary**

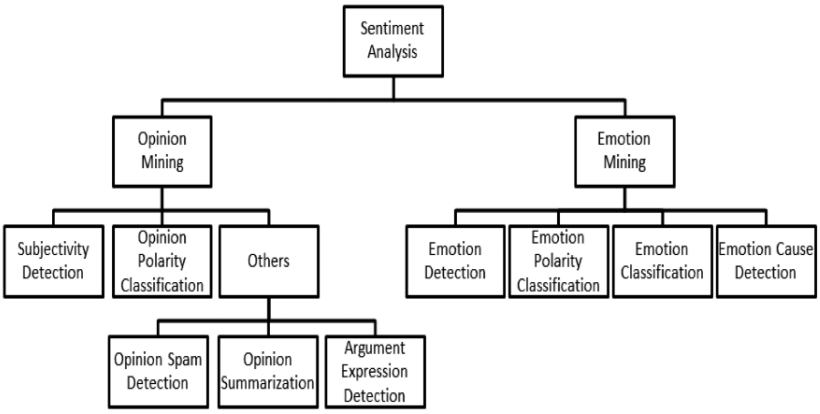

Figure 3.1: Taxonomy of sentiment analysis (Yadollahi et al., 2017, 25.2)

classification as proposed in the earliest studies implying that any input to be classified is either `positive` or `negative`. However, it is quite easy to come up with an input, such as the phrase *today is Wednesday* which lacks any word with a clear semantic orientation. As hinted at earlier, instances of this kind are assigned a `neutral` label in some studies whenever there is no explicit or implicit polarity-wise contextualization. One such polarity affecting context could be a biographical text about an individual whose most or least favorite day happens to be a Wednesday. Besides a **ternary classification**, some studies conduct a **quaternary** one which includes a class for instances characterized by both negative and positive aspects. This class is commonly called `mixed` (Davoodi and Mezei 2022, 219; Mæhlum et al. 2024, 10). In summary, this means that most approaches (implicitly) classify a given input with some kind of function that assigns one of these two to four polarity labels. Another line of work prefers to use **ordinal scales**, such as star ratings, instead (Elkins, 2022, 24). Finally, there are also approaches which rely on continuous polarity intervals sometimes called **valence** (see Section 3.1.2).

Target granularities

Sentiment analysis can be applied to different levels of text granularity as illustrated by Figure 3.2:

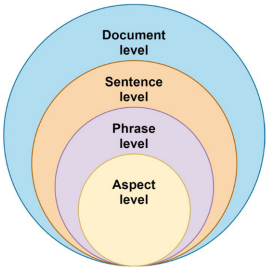

Figure 3.2: Sentiment analysis levels (Wankhade et al., 2022, 5734)

The most coarse-grained approach is the **document level** where only one sentiment label or polarity score is attributed to a whole text document. According to Wankhade et al. (2022, 5734), this level "is not used a lot". This might be due to the fact that polarity differences tend to average out over longer spans of text. On the other end of the scale, there is the **aspect level** which is the most fine-grained kind of sentiment analysis and the only level accounting for the sentiment target (Zhang et al., 2018). However, said level adds the difficulty of identifying the target. This is especially challenging in mixed cases where multiple targets need to be distinguished from each other. Most approaches operate on the intermediate **phrase** or **sentence level** (Wankhade et al., 2022, 5734–5735). Still, all levels are permeable in the sense that an aggregation to a higher more coarse-grained level is always possible.

3.1.2 Computational methods and resources

Computational approaches towards a sentiment classification typically fall into a continuum that ranges from dictionary lookups over more "traditional" machine learning to deep learning systems while also including hybrid approaches (Ali et al., 2025, 4007–4016). The most relevant system types will be portrayed in the following with a focus on the German language and digital humanities research.

Dictionary-based approaches to sentiment analysis

Kim and Klinger (2021, 11) noted that dictionary-based systems are more common in digital humanities than in computational linguistics presuming that this might be associated with transparency issues. Similar considerations were observed by Van Atteveldt et al. (2021, 134) in computational social sciences. It is therefore not surprising that concerns of this kind were also raised in the work closest to this one. More specifically, McGuire (2021, 54) and Dennis-Henderson et al. (2020, 93–94) both stated that they analyzed biographical datasets with sentiment dictionaries in order to benefit from their transparent algorithmic decisions.

With respect to the German language, the survey of Kern et al. (2021) lists 15 dictionaries and the one of Fehle et al. (2021) 19 each of which are readily available. These resources are either compiled by human raters, such as BAWL-R (Võ et al., 2009), or augmented from a seed list in a semi-supervised machine learning setting, such as SentiWS (Remus et al., 2010).

Regarding a manual compilation, Koto et al. (2024, 299) differentiated between a **"direct annotation"** on the one hand and a **"best-worst scaling"** on the other. In the first case, annotators assign a polarity score or sentiment label to each word form directly whereas they iteratively select the most positive/negative word form from a subset in the second case.

It is also worth mentioning that German sentiment dictionaries have different sizes as AffectiveNorms (Köper and Schulte im Walde, 2016), for instance, comprises more than 350,000 word forms but LANG (Kanske and Kotz, 2010) only 1000. Both sizes are edge cases since "German resources contain usually less than 3000 words" according to Köper and Schulte im Walde (2016, 2595). In the case of an inflectional language like German, some of the size differences between dictionaries arise because some list only lemma forms while others include inflected ones, too. When using a dictionary of the former kind, it might be worth considering further preprocessing steps like a lemmatization, even more so as this may lead to performance gains. In the survey of Fehle et al. (2021, 102), for instance, the average F1 score increased from .45 to .50 after lemmatization even though both values are quite similar and far from perfect. Another possibility for improving the performance is the inclusion of contextual rules in order to account for the influence of valence shifters, such as the negator *not*. One system of this kind is VADER (Hutto and Gilbert, 2014) which has been adapted to German by Tymann et al. (2019).

While the size of a dictionary is a rather coarse measure for an a priori assessment of the applicability, it can be complemented by a **coverage analysis**. In this way, Koncar et al. (2020, 7) observed that common dictionaries covered the least amount of word forms in the case of 18th century Portuguese periodicals (= 25 %) and the second least (= 36 %) in the case of German ones. In another study of textual sources from a similar time frame, Herrmann and Grisot (2022, 168) also observed rather low coverages although there are differences between different dictionaries (see Table 3.1). As a matter of fact, most of the dictionaries they evaluated covered only 3–17 % of all words in their corpus. One notable exception was SentiArt which tagged more than 80 % but is rather an emotion than a sentiment dictionary and hence out of scope here. BAWL-R and LANG, by contrast, just use the synonym *valence* for *polarity* so that they are actually both sentiment and emotion dictionaries. The rather low coverages of most systems may relate to domain particularities, language change, and graphematic differences. Therefore, assessing the coverage is useful as a (too) low one will result in most instances being labeled as `neutral` which impairs downstream analysis. Yet, it is difficult to know a priori if a given coverage is "too low" since one can imagine texts which are just not characterized by many sentiment-bearing words.

Lexicon	Coverage	Dimensions
Adu	.04	joy, fear, sadness, surprise, disgust, anger, depression, love
BAWL-R	.32	**valence**, arousal, imageability
Germanlex	.17	**polarity**
Klinger	.09	joy, fear, sadness, surprise, disgust, anger
LANG	.07	**valence**, arousal, concreteness
Plutchik	.03	joy, fear, sadness, surprise, disgust, anger
SentiArt	.82	joy, fear, sadness, surprise, disgust, anger, AAZ
SentiWS	.13	**polarity**

Table 3.1: Dimensions listed in common German sentiment and emotion dictionaries (Sentiment equivalents marked in bold here) and their coverage when applied to 76 Swiss prose texts written between 1854 and 1930 (Herrmann and Grisot, 2022, 168)

"Traditional" machine learning approaches to sentiment analysis

Concerning "traditional" machine learning approaches, especially **SVMs with a linear kernel** have been shown to be among the most performant sentiment classification systems in many surveys (Waltinger, 2010a; Singh and Singh, 2021; Ali et al., 2025). Wankhade et al. (2022, 5753) even claimed that an SVM is the "[m]ost popular algorithm" for sentiment analysis although this approach comes at the cost of a "[l]ong training time for large datasets". One reason for the popularity could be the fact that already Pang et al. (2002) suggested the use of SVMs. Nevertheless, SVMs have never been as widespread as dictionaries and more recently transformer-based models in digital humanities projects conducting sentiment analyses. A few notable exceptions are the happy ending classification by Zehe et al. (2016) as well as three sentence-wise sentiment classifications (Zehe et al., 2017; Du and Mellmann, 2019; Rebora et al., 2022). The first three studies deal with German novels and the latter one with book reviews. All of them include a sentiment dictionary lookup to gain some of the features for their respective SVM classifiers. Hence, said approaches are in fact hybrid ones with regard to the systematization of Ali et al. (2025, 4007) and substantiate the prevalence of lexicon-based ones in digital humanities even further.

Interestingly, Du and Mellmann (2019, 8) motivated their model selection based on the suggestion to use SVMs when there are less than 100,000 manually annotated instances available. In addition, they compared various vectorization settings and found that a TF-IDF weighting is beneficial (Du and Mellmann, 2019, 12). The same is true for the study of Davoodi and Mezei (2022, 226) which deals with the common sentiment application domain of customer reviews.

Deep learning approaches to sentiment analysis

Ali et al. (2025, 3674) identified notable diachronic trends during their literature review of 60 research papers written between 2013 and 2023: Whereas lexical approaches to sentiment analysis played a major role up until 2015 and "traditional" machine learning ranked among the top two most used ones until 2018, they both ceased to appear in the papers they analyzed from 2019 onwards. This turning point may be explained with the introduction of the transformer model BERT by Devlin et al. (2019). In the case of German sentiment analysis, Corvonato (2021, 40) showed that all five BERT-based models she evaluated outperformed the previously best model for the document-level ternary sentiment classification of the GermEval-17 shared task (Wojatzki et al., 2018). This way she was able to increase the "state-of-the-art" micro F1 score to .80. With regard to German book reviews, Rebora et al. (2022, 355–356) reached a macro F1 score of .85 with their fine-tuned BERT model. Their second best approach was an SVM trained on sentiment dictionary features and outperformed by a remarkable 21 % point margin. Schmidt et al. (2021) focused on emotion analyses of 18th and 19th century German plays but included a binary sentiment classification, too. They also found that BERT-based models yield F1 scores in the range of .79–.82 outperforming their SVM (.59–.64) and lexicon approaches (.45–59).

Concerning off-the-shelf BERT models fine-tuned for German sentiment, it is worth mentioning that the performance of german_sentiment_bert (Guhr et al., 2020) was found to be quite sensitive to the target domain. Klähn et al. (2025, 10), for instance, showed that it yielded an F1 score of .81 when applied to customer reviews but only .33 in the case of humanities datasets. Hence, the intended target domain is not only important for dictionary-based systems.

While most related work discussed so far relied on **supervised** approaches to machine/deep learning, a growing interest in **unsupervised** ones in general and **zero-shot settings** in particular is recently observable. As sentiment analysis is a common task, it is not surprising that generic models, such as BERT fine-tuned for the NLI task, have already been assessed in this manner. For instance, Laurer et al. (2024, 92) showed that such a zero-shot setting outperformed SVMs with respect to a binary sentiment dataset. Similarly, Klähn et al. (2025, 4) compared that approach with sentiment lexicons, off-the-shelf BERT models fine-tuned for German sentiment, and prompt-based LLMs. They found that zero-shot approaches in general and the LLM *Llama3.1* in its 80B version in particular are able to outperform smaller BERT models.

3.1.3 Typical downstream tasks in digital humanities

While some digital humanities studies, such as the one of Klähn et al. (2025), focus on system comparisons, many more analyze sentiment information with a different downstream task in mind. One line of research involves **social network analyses** like the ones Nalisnick and Baird (2013) conducted for Shakespeare's plays. In addition, they investigated **genre differences** and found that Shakespeare's comedies are more positive than his tragedies. A different line of research looks into **diachronic trends**. One such example is the analysis of multilingual periodicals from the Age of Enlightenment (Koncar et al., 2020). Said downstream task builds up on one of the earliest approaches to sentiment analysis (see Section 3.1.1). However, Koncar et al. (2020, 12) grouped their time series by language and topics instead of whole documents. In the case of biographical sources, Dennis-Henderson et al. (2020, 97) investigated topic differences over time as well.

Faull and McGuire (2022, 143–144) focused their analyses of English Moravian memoirs on gender differences in **trajectories over the narrative time**. These were generated with the Syuzhet package Jockers (2015) introduced in a blog post. Said approach was inspired by Vonnegut's rejected master thesis in Anthropology. According to his later biographical remarks, Vonnegut (1981, 285) wanted to prove that there are only a few "basic shapes" most stories follow when looking at the "fortune" of the protagonist. Especially after Jocker's blog post, both works led to a rise of various similar studies in the field of computational literary studies (Rebora, 2023, 36). One example is the work of Reagan et al. (2016, 5) who identified six **"emotional arcs"** in 1327 books from Project Gutenberg by clustering. Figure 3.3 displays the top 3 modes derived from a Singular Value (SV) Decomposition of the clustered sentiment time series. They call the rising positive SV 1 "Rags to riches" and its falling negative counterpart "Riches to rags", SV 2 "Man in a hole" (falling-rising) or "Icarus" (rising-falling), and SV 3 "Cinderella" (rising-falling-rising) or "Oedipus" (falling-rising-falling).

A similar categorization of "basic plot shapes" derived from sentiment analyses were proposed by Brown and Tu (2020). Their categorization focuses on polarity shifts in the beginnings and endings of stories. Moreover, it is based on a slightly different underlying sentiment concept – namely *happiness*. This leads to four patterns (see Figure 3.4) that are similar to the aforementioned ones. More specifically, a "Quest" corresponds to the "Cinderella" story of Reagan et al. (2016), and a "Looser's Tale" to "Icarus". The "Man in a hole" shape, by contrast, has the same name in both studies, whereas a "Tragedy" corresponds to "Oedipus".

Furthermore, Brown and Tu (2020, 269) added the **harmonicity** concept. It accounts for their assumption that patterns, which only differ from each other due to different numbers of shifts in intermediary parts of a text, are actually quite similar. They would hence call all three patterns depicted in Figure 3.5 "Quest" with different "harmonic expansions". This concept allows them to relate most possible sentiment trajectories over narrative time to these four "basic shapes". Although they do not make it explicit, there likely is a zeroth Harmonic, too, that accounts for a constant rise or fall since these are the only two patterns they lack compared to Reagan et al. (2016).

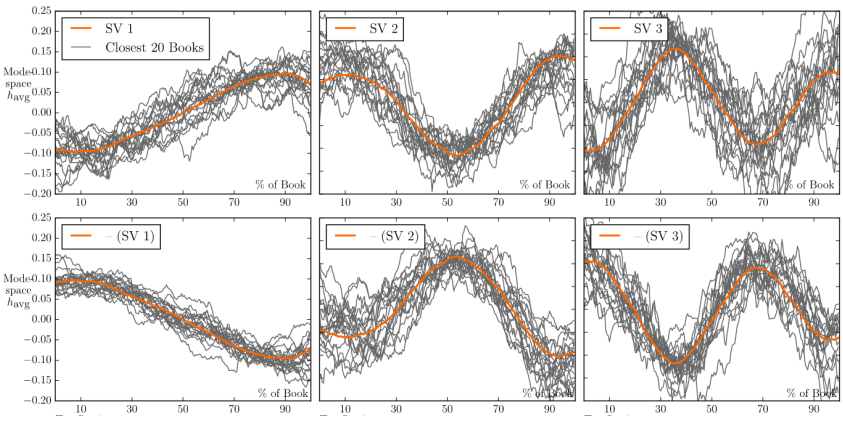

Figure 3.3: Top three Singular Value (SV) Decomposition modes and their negations of 1327 books from Project Gutenberg (based on Reagan et al. 2016, 7)

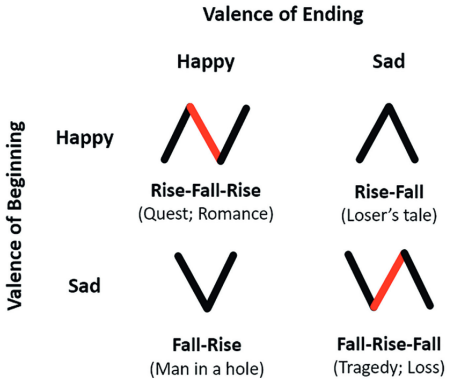

Figure 3.4: A 2 x 2 crossing between the valence of the beginning (the initial shift) and the ending (the final shift) of a story (Brown and Tu 2020, 267)

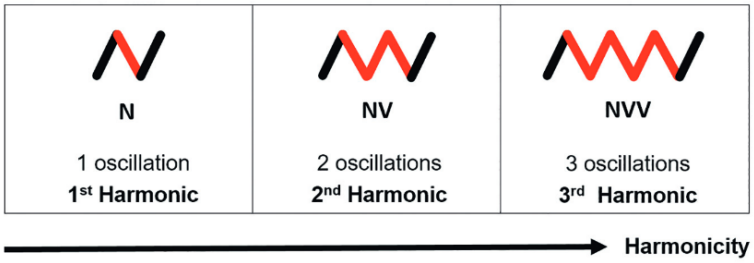

Figure 3.5: The "Quest" shape and its first two harmonic expansions (Brown and Tu 2020, 269)

3.1.4 Typical issues and methodological critique

Although sentiment analyses have become part of the methodological canon of various disciplines as illustrated by the examples above, there are some issues each study needs to resolve. One of the most apparent ones is **domain-sensitivity** and was already hinted at with the exemplary phrase *today is Wednesday* in Section 3.1.1. In the case of said phrase the aforementioned specific "domain" of individuals known to treasure or hate Wednesdays for some reason or another may seem fictitious. Yet, certain words and phrases have been shown to bear different sentiment polarities in different textual domains (Hamilton et al., 2016). Hence, it is likely the case that current approaches struggle with humanities datasets because the former tend to be optimized for other domains, such as customer reviews and social media data (see Section 3.1.2) This is especially true for the German language as Guhr et al. (2020, 1627) stated that "all available German [sentiment] datasets reflect social media language rather than general-purpose German". In the case of (English) Moravian

memoirs, Faull and McGuire (2022, 133–134) noted issues with domain-specific language usage when their dictionary attributed different polarities to the words *sister* and *brother*. No matter if this disparity may be related to gender biases or not, it is problematic because Moravians used these terms as generic ways to refer to other members of their denomination (see Section 4.1.2). More generally, these observations are in line with the claim of Fehle et al. (2021, 94) that "lexicons that are designed for specific corpora or domains perform best on these corpora". This applies to fine-tuned BERT models, too, as shown by Klähn et al. (2025). To mitigate said issue, Elkins (2022, 46) suggested "comparing different families of models [...] utilizing the wisdom of crowds approach in which more perspectives are usually better."

Besides domain-specific peculiarities, Ali et al. (2025, 4024–4025) considered the handling of **ironic** or **sarcastic** expressions as one key challenge. These cases are characterized by a disparity between what is said and what is meant. To this list, one could add **euphemistic** expressions because they make it even more important to differentiate between an expressed sentiment polarity on the one hand and an implied one on the other. Whereas these issues impact manual and automatic annotations likewise, problems related to the influence of degree adverbs and intensifiers (Wankhade et al., 2022, 5768) are even more prominent in the latter case. The same is true for the observations that some sentiment datasets tend to be skewed towards the `positive` class (Ali et al., 2025, 4025) while others include more samples with the label `negative` (Schmidt et al., 2018, 50). This implies that models trained on existing datasets may require an explicit debiasing.

Ali et al. (2025, 4025) and Wankhade et al. (2022, 5769) both noted that **informal language** is difficult to analyze because it contains abbreviations, acronyms, typos, and non-standard grammatical variation when referring to peculiarities of social media content. Similar problems can be expected to arise from **historical language**. This affects not only word forms but also punctuation which is one reason why a sentence segmentation is more difficult for historical language stages than for contemporary ones (McGuire, 2021, 81). Finally, the influence of **language change** needs to be mentioned as this can also affect sentiment polarity. Hamilton et al. (2016, 596), for instance, showed that the adjective *terrific* flipped its polarity to `positive` between the 1950s and 1970s after consistently bearing negative sentiment for at least one century.

Methodological critique

One line of methodological critique is related to the observation that sentiment analyses are often conducted with multiple quite distinct concepts in mind. This is problematic as shown in a survey of 60 research papers (Venkit et al., 2023, 13747–13748) where, as matter of fact, only "**31** [...] explain[ed] the employed framework, and **2** [...] explicitly define[d] sentiment." In a similar vein, Van Atteveldt et al. (2021, 136) recommended formalizing both the conceptualization and operationalization. In this context, they also highlighted the importance of a manually annotated sample from the target domain in order to assess the validity of the system(s) used. Further validity issues arise from the observation that commonly used dictionaries tend to underperform compared to systems like SVMs (Singh and Singh, 2021) or transformers (Corvonato, 2021).

Venkit et al. (2023, 13749) stated that sentiment models trained on large scale corpora may potentially contain harmful biases with respect to age, gender, and racial stereotypes. Such ethical risks may also affect personal names because "according to the distributional hypothesis, the embeddings of [...] names will be close to the words with explicit sentiment with which they co-occur" (Hube et al., 2020, 259). However, respective debiasing efforts are far from being common practice. They may also be more difficult to achieve in the case of systems characterized by a rather opaque decision-making process, such as deep learning approaches. In the case of dictionaries, by contrast, a debiasing is easier to accomplish as it only requires to change the respective values.

Regarding downstream tasks, it should be noted that the Syuzhet package is or at least was quite popular in digital humanities projects due to its ability to generate trajectories, but is also not undisputed. The main issue relates to its smoothing approach as it initially used a Fourier transformation which can lead to ringing effects if a given text does not start and end with neutral sentiment (Elkins, 2022, 11). This is a problem when trying to identify and compare "basic shapes" as proposed by Reagan et al. (2016) and Brown and Tu (2020).

Further criticism from the field of computational literary studies concerns the custom of calling sentiment trajectories "plot arcs". Rebora (2023, 7), for instance, stated that this term contradicts narratological theory where *sentiment* is seen as an affect category and "affect is nothing but a device used to engage the reader, with just secondary and indirect consequences on the structure of the narrative". Therefore, the critique basically relates to (i) terminological questions and (ii) the methodological question of whether sentiment should or should not be used as a proxy for a concept like *plot* as a whole. It does, however, not relate to the methodological question of whether one should or should not try to find sentiment polarity patterns in texts.

3.2 Subdivision of the life course

Whereas sentiment (polarity) assesses the way a narrative is depicted, one can also account for the structure of a narrative. In the case of dramatic texts, for instance, scenes would be a fitting narrative unit, whereas prosaic works can typically be subdivided into chapters or paragraphs. However, due to the diversity of biographies (see Section 2.1), these sources lack a clear-cut counterpart. Still, Werner (2022, 396) noted that biographical sources tend to cover topical sections dedicated to *birth*, *genealogy*, *baptism*, *childhood*, *school life*, *work life*, *marriage*, *child birth*, *virtues and shortcomings*, *illness*, and *death* from the 17th century onwards. As a matter of fact, similar sections can also be found in the data investigated here (see Table 4.1). Nevertheless, this kind of subdivision is rather subgenre-specific as it disregards the possibility that some biographies may follow a different narrative. It also combines events like *birth* or *marriage* with periods of time, such as *childhood* or *adulthood*. This is why a more general approach applicable to any biographical source is followed here. More specifically, less emphasis is put on the exact order in which sections occur in a narrative and more emphasis on the narrated time since all biographies describe (parts of) what is called the **life course** of an individual. To this end, Section 3.2.1 offers definitions of the term *life course* as well as related concepts by taking differences between various scholarly traditions into account. Afterwards, Section 3.2.2 shows how these concepts can be operationalized computationally before Section 3.2.3 summarizes issues that may arise from operationalizations.

3.2.1 Life course, life stages and similar concepts

An often cited definition of the term *life course* is the following:

(3.3) *a sequence of socially defined events and roles that the individual enacts over time (Giele and Elder, 1998, 22)*

This definition highlights the temporal aspect and shows that the focus of inquiry depends on the field of research which clearly is sociological. However, life courses have been studied in many more fields of research ranging from biology (Geifman et al., 2013; Mousley et al., 2025) to history (ter Braake and Fokkens, 2015; Meister, 2018). This is why Bernardi et al. (2019) opted for a multi-dimensional view of the life course that comprises the three dimensions (i) **time**, (ii) **life domain**, and (iii) **level** (see Figure 3.6). The third dimension will be henceforth sidelined because biographical records typically describe the intermediary level of **individual actions**. This does not mean that, for instance, the lower **inner-individual** level, which biological and psychological research focuses on (Bernardi et al., 2019, 5–6), is deemed less important, it is only less accessible on the base of these sources. Still, it is likely that the biographies under investigation hint at some effects, such as developmental differences between children and adults. In a similar vein, the **supra-individual** level encompasses sociocultural particularities of the subjects whose biographies are analyzed that may be of interest but are difficult to assess based on historical sources alone. Still, a few effects may be observable in the comparative content analysis portrayed in Section 9.4. When mainly considering the mesolevel, while acknowledging interdependencies with the other two levels, the time dimension equals **life-course trajectories**. These are the focal point of the remainder of this section, whereas the domain dimension will be explored further in Section 3.3.1.

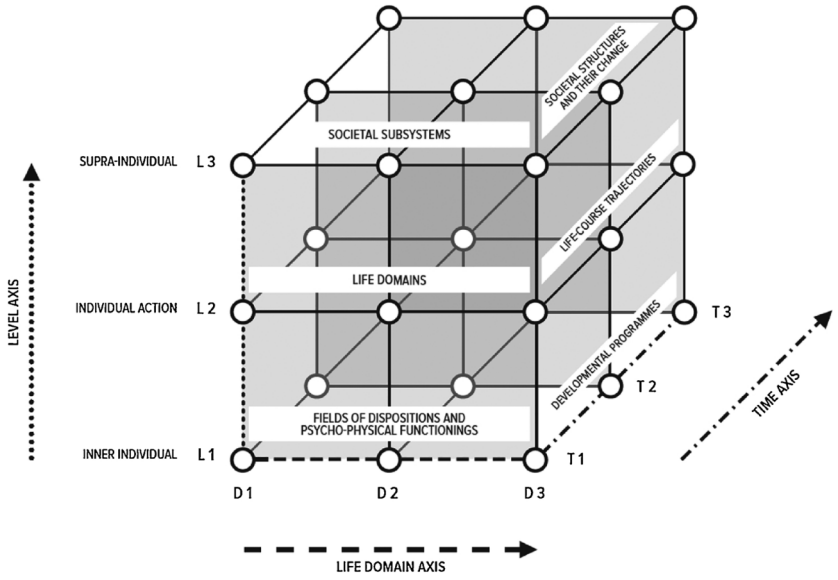

Figure 3.6: Life course cube (Bernardi et al., 2019, 4)

The temporal dimension is mostly linear and typically involves a so called **path dependency** – i.e. "the universe of possible future pathways is restricted due to the causal impact of the previous biography and decisions" in the words of Bernardi et al. (2019, 3). This has implications for biographies as a genre since the typical goal of a retrospective sense-making may require non-linearities in the overall narrative, such as prolepses and analepses. In addition, it is worth pointing out that the interdependencies between the different levels have led to multiple ways in which the individual life time can be subdivided in theory and practice. This is even more true given that there are also different cultural conceptions of time, as pointed out by Titzmann (1996, 156). Yet, most subdivisions can be categorized into (i) ones based on **turning points**, and (ii) ones that group people with a similar age into **life stages**.

Turning points

According to Bernardi et al. (2019, 4), a turning point "reflects a radical deviation or disruption in the trajectory an individual has been on or from one that was personally or socially expected in the future." Hence, it refers to a central event an individual experiences and is likely incorporated into a biographical source as the transition between two passages of life. Such an event can be categorized with regard to (i) the "roles affected", (ii) the "source or cause", (iii) the "timing", and (iv) the "ultimate consequences" (Clausen, 1998, 204). Anthropological research has shown that especially role or status changes are often ritualized in many societies. Typical instances of these so called **rites of passage** are "birth, childhood, social puberty, betrothal, marriage, pregnancy, fatherhood, initiation into religious societies, and funerals" (Van Gennep, 1960, 3). These events are of particular interest for this thesis because rites of passage have also become *grundlegend* 'fundamental' for religious studies dedicated to biographical sources as pointed out by Dormeyer (2022, 523). A hermeneutic approach in this field would therefore most certainly also take the other three mentioned parameters (= source or cause, timing, and ultimate consequences) into account. However, they are less important when assessing turning points with regard to their narrative structuring potential. Yet, as the second last one is relatable to age (groups), it is not surprising that some turning points bridge the gap to this second way of subdividing the life course. This is not only true for supra-individual sociocultural ones but also for the inner-individual level. A notable example from the field of cognitive neuroscience is a large-scale study in which "five major epochs of topological development, each with distinctive age-related changes in topology" were identified when investigating more than 4000 individuals (Mousley et al., 2025, 1).

Life stages

The earliest (known) work dedicated to the categorization of the life course by age into discrete units, which are commonly called **life stages**, dates back to the Greek and Roman antiquity. More specifically, Lievegoed (1981, 33–34) stated that Ancient Greek authors divided the life into ten ἑβδομάδες 'seven year spans', whereas Roman ones grouped individuals by age into the five life stages (i) *pueritia* 'childhood' (0–15 years), (ii) *adolescentia* 'adolescence' (15–25), (iii) *iuventus* 'youth' (25–40), (iv) *virilitas* 'manhood' (40–55), and (v) *senectus* 'aged' (> 50). In early modern times, seven to ten age groups were common but subsequently reduced to four in the 18th century albeit that both the respective **age ranges** as well as group labels vary considerably (Titzmann, 1996, 159–160). The latter is also true for other common subdivisions in different disciplines as described in the following. This is why only age ranges are depicted in the summarizing Figure 3.7.

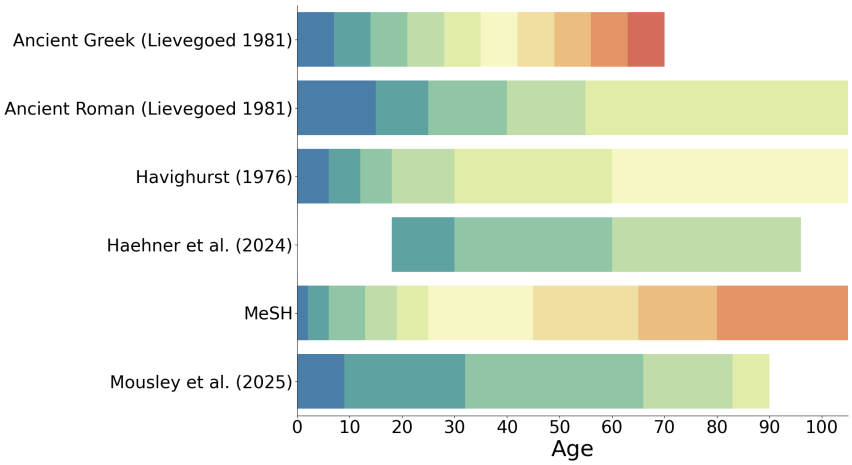

Figure 3.7: Overview over different age-based subdivisions of the life course

In modern sociological and psychological studies, subdivisions are either operationalized "etically ([based on] arbitrary, but equal age spans[,] such as decades) or emically (related to life stage[s], such as adolescence etc.)" according to Nguyen et al. (2011, 116). An emical subdivision is typically also hierarchical with a distinction between (i) **"childhood"**, (ii) **"adulthood"**, and (iii) **"old age"** on the highest level. Further subdivisions may include splitting (i) into *"early childhood"*, *"youth and adolescence"*, and *"post-adolescence"*, (ii) into *"early adulthood"*, *"midlife "*, and *"old age"*, and (iii) into *"young-old"* and *"old-old"* (Settersten and Mayer, 1997, 249). However, the exact terminology varies between studies which most probably explains why the term *old age* appears in two different positions of said hierarchy. In addition, the respective life stage labels are usually not defined with regard to exact age ranges in order to account for individual differences (Thomae, 2002, 22–23). Yet, there are exceptions. Havighurst (1976), for instance, noted that the six periods (i) "Infancy and Early Childhood", (ii) "Middle Childhood" (6–12 years), (iii) "Adolescence" (12–18), (iv) "Early Adulthood" (18–30), (v) "Middle Age" (30–60), and (vi) "Later Maturity" are characterized by different *developmental tasks*. It is worth noting that Havighurst (1976, e.g. 19) qualified all age ranges with the word "about" and does not provide ones for the initial and final period even though they are inferable from the others. In a more recent psychological study, Haehner et al. (2024, 5) seemingly followed this subdivision although they (i) investigated people who were 18 to 95 years old so that the first three age groups were irrelevant, and (ii) called the latter three "young adulthood (18–30 years[...]), middle adulthood (30–60 years[...]), and old adulthood (> 60 years[...])".

In the life sciences, it is more common to define age groups by age ranges and an often cited classification is the one of the *Medical Subject Headings (MeSH) thesaurus*[4]. It differentiates between (i) "Infant" (0–2 years), (ii) "Child, Preschool" (2–5), (iii) "Child" (6–12), (iv)"Adolescent" (13–18), (iv) "Young Adult" (19–24), (v) "Middle Aged" (25–64), and (iv) "Aged" (> 64) (Geifman

[4] https://www.ncbi.nlm.nih.gov/mesh/68009273 (Last accessed on February 18, 2026)

et al., 2013, 2359). However, there still exist different subdivisions. The aforementioned neuroscientific study, for instance, proposed the age ranges (i) 0–9, (ii) 9–32, (iii) 32–66, (iv) 66–83, and (v) 83–90 based on the topological differences they identified without attributing specific labels (Mousley et al., 2025, 6). In summary, it is noteworthy that most subdivisions comprise longer periods during adulthood than childhood. This can be attributed to the observation that a specific age has less impact on the lives of adults than younger ones. This led Giele (1980) to call this life stage the "[t]ranscendence of [a]ge".

3.2.2 Computational methods and resources

To the best of my knowledge, there have not yet been any previous computational attempts to operationalize the identification of life stages in (historical) biographies. However, social media research involves the somewhat related age identification task. This author profiling subtask aims at the prediction of the age (group) of a user. Due to the difficulty of collecting large-scale datasets annotated with the actual age a user has in the content, the target label is often approximated by the age at which she or he uploaded the respective post (Shrestha et al., 2016; Hsieh et al., 2018; Lupo et al., 2024). Another age approximation relates to the actual task operationalization which is usually a single age group label classification. However, a few studies, such as the one of Nguyen et al. (2011), use a regression in order to predict the exact age instead. System-wise, Wang et al. (2019), for instance, implemented said task in a multi-modal fashion that leverages the availability of profile images. Hence, this line of research differs not only with respect to the totality from an analysis of historical biographies but also with regard to media types.

3.2.3 Typical issues and methodological critique

The practice of subdividing the life course into life stages is not undisputed (Lievegoed, 1981, 31) with the main issue being the fact that developmental differences between humans exist (Thomae, 2002, 22–23). Thus, every subcategorization may appear arbitrary to some extent as it disregards said differences, and may lead to margins between adjacent categories that are difficult to assess. Yet, subdivisions remain universal in the sense that they are not only used in sociological studies but also backed by anthropological and neuroscientific ones (see Section 3.2.1).

Another issue arises from the observation that there are also cultural and diachronic differences (Giele and Elder, 1998, 9–10) so that a modern day categorization may seem anachronistic when applied to 18th century people. One such difference is the shorter life expectancy which was only 35.1 years on average in the Americas and 36.3 in Europe in 1850 (Riley, 2005, 538). Therefore, biographical records from that time are less likely to include sections about later life stages in theory. However, most biographical subjects investigated in this thesis actually outlived said life expectancies (see Figure 4.4) so that this is a smaller issue here.

Regarding alternative approaches towards predicting age, it should be added that they also have disadvantages. For instance, both Nguyen et al. (2014, 1957) and Flekova et al. (2016, 851) showed that labeling the exact age based on textual features alone is even difficult for human annotators. Similarly, a focus on turning points disregards the fact that "transitions from one state to another may be gradual, not abrupt, more of a process than an event" in the words of Karweit and Kertzer (1998, 94). Furthermore, there is the more practical issue that not all biographies are required to be structured based on events of this kind alone.

3.3 Biographical event extraction and categorization

Despite being less apt for a *structural* segmentation, events still remain important *topical* building blocks of life narratives which are worth investigating. Karweit and Kertzer (1998, 93), for instance, viewed "life histories as a collection of interconnected event histories in separate domains". In a similar vein, Bamman and Smith (2014) showed how life events connect individuals of the same age and/or period of time. Life events hence account for the inner- and supra-individual levels of the life course cube (see Figure 3.6), and complement an age-based subdivision. Event models are also

relevant in (computational) literary studies so that there are different theoretical views on how to approach this concept. These are contrasted to each other in Section 3.3.1 before Section 3.3.2 gives an overview over computational implementations. Section 3.3.3 builds up on this by summarizing which typical issues computational operationalizations have encountered in related work.

3.3.1 Biographical events and similar concepts

"Events can be considered the smallest units that make up a narrative" in the words of Solissa et al. (2025, 1). This is why they play a major role not only in biographical research but also in literary studies. A common definition of the term *event* in this field of research is the following:

(3.4) *words or phrases (typically verbs, sometimes nouns) that clearly signal, on the linguistic surface, the existence of a specific action, activity, or change of state*
(Rehm et al., 2017, 43)

The fact that this definition relates events with verbs is rooted in linguistic typology where events are sometimes seen as an aspectual feature of non-stative predicates. One such predication typology is proposed by Mourelatos (1978, 422–424) and depicted in Figure 3.8. In this typology, which is based on binary contrasts, **states** are distinguished from what he calls **occurrences** (or **actions**) on the highest level. The latter encompasses **processes** besides **events** which, in turn, can be further subdivided into **developments** and **punctual occurrences**.

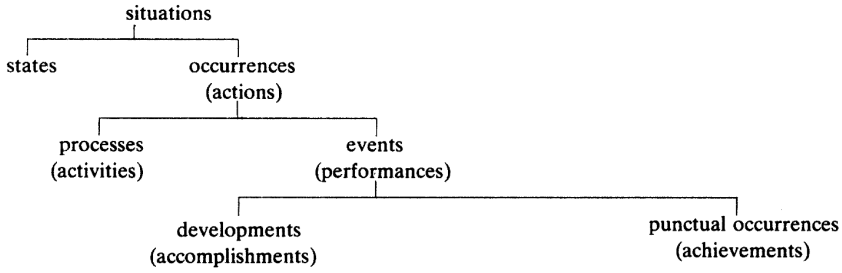

Figure 3.8: Events in a verb predication typology (Mourelatos, 1978, 423)

Narrative event types

The top level dichotomy between states and occurrences/actions implies that the latter involve a **change of state**. Said concept is also prominent in the case of narratological event definitions (see Definition (3.7)). In addition, there is typically a two-fold distinction between type I and II events in this field of research which Hühn (2014, 159) defined as follows:

(3.5) *A type I event is any change of state explicitly or implicitly represented in a text.*

(3.6) *A change of state qualifies as a type II event if it is accredited – in an interpretive, context-dependent decision – with certain features such as relevance, unexpectedness, and unusualness.*

In their review of narrative theory, Gius and Vauth (2022, 4) identified an "anthropomorphic agent" as well as high degrees of intentionality as the two most widespread additional type II requirements. Moreover, this second event type depends on "cultural and social historical contexts" so that "[d]etermining eventfulness is therefore a hermeneutic process" that allows for gradations according to Hühn (2014, 171). In the context of biographical sources, it can be expected that there is a focus on major life events like the aforementioned turning points (see Section 3.2.1) which have some ties to the definition of a type II event. Furthermore, Hühn (2014, 165) claimed that, "[f]or the individuals involved and, seen from their perspective, changes defined as rites of passage possess the features of a type II event" which also links to aspects of the previous section.

Another more recent narratological approach towards subcategorizing events is the one Vauth et al. (2021, 336) derived from different earlier categorizations. These seemingly include linguistic typology as implied by their focus on verbal phrases. More specifically, they distinguished between four event types which they defined in the following manner:

(3.7) *The event type* ***<change_of_state>*** *covers physical and mental state changes of animate and inanimate entities. The change of state needs to be expressed in a single verbal phrase.*

(3.8) *The event type* ***<process_event>*** *covers actions and happenings that do not lead to a change of state, such as processes of moving, talking, thinking and feeling.*

(3.9) *The event type* ***<stative_event>*** *covers all verbal phrases that refer to the state of an animate or inanimate entity. This also includes physical and mental states. Unlike changes of state and processes, stative events do never refer to temporal processes.*

(3.10) *The event type* ***<non_event>*** *finally covers verbal phrases that do not refer to a fact in the narrated world. The most common variants of non-events are questions or modalized and generic statements[.]*

These four event types differ with regard to their eventfullness in descending order. In this context, it should be stressed that not only the linguistic predication system of Mourelatos (1978) but also other subcategorizations in the field of computational literary studies tend to not consider the less eventful classes instances of an event. In a similar vein, events are typically required to be real and not, for instance, imagined or hypothetical as is the case for the classes `stative_event` and `non_event` (Solissa et al., 2025, 5–6).

Major life events

In contrast to literary texts, the identification of events in biographies can be deemed less important than their categorization because these sources mainly deal with events (see Section 2.1.1). More specifically, life narratives are typically composed around so called **major life events**. Some events of this type are somewhat universal and correlate with certain ages or age groups while others do not. To give some examples, Brim (1980, 153) categorized 37 exemplary life events by (i) the amount of experiencers, (ii) the probability of their occurrence, and (iii) their correlation with age. These are depicted in Table 3.2 and range from very common ones like *starting to work* to very specific ones, such as *winning a lottery*. Interestingly, there is also a correlation with sentiment polarity. For instance, Haehner et al. (2024, 2) noted that "positive life events like graduation or starting a romantic relationship are expected to occur primarily in young adulthood [...], whereas there are no clear age-graded expectations for negative events". More generally, these exemplary major life events also hint at the fact that – at least in theory – there is an unlimited amount of events which may be categorized this way since basically any event can have a lasting impact on the life of one individual or another. This is why Karweit and Kertzer (1998, 93) opted for grouping such events by the respective "[d]omain or life areas such as education, occupation, and family formation" they concern as "it is extremely useful to consider them as separate areas for measurement and data organization purposes".

Event domains

While life event domains can be categorized based on a priori groupings of common life events, such groupings can also be refined empirically by considering events individuals deemed important. This approach was recently followed in a large scale study where a "nationally representative sample of German inhabitants" was asked to name the "most important major life event they had experienced in the last 12 months" which yielded 1044 free-text answers (Haehner et al., 2024, 4–5). The resulting hierarchy comprised 50 fine-grained event domains and 13 more coarse-grained ones the latter of which are defined as follows in the respective coding instructions[5]:

[5] https://osf.io/x3ucv/files/93u5j (Last accessed on February 18, 2026)

Exp.	Prob.	Strong Correlation with Age	Weak Correlation with Age
Many	**High**	marriage	
		starting to work	death of a father
		retirement	death of a husband
		entering school	male testosterone decline
		woman giving birth to first child	"topping out" in work career
		Bar Mitzvah	children's marriages
		first walking	accidental pregnancy
		heart attack	
		birth of sibling	
	Low	military service draft	war
		polio epidemic	Great Depression
			plague
			earthquake
			migration from South
Few	**High**	heirs coming into a large estate	son succeeding father in family business
		accession to empty throne at 18	
	Low	spinal bifida	loss of limb in auto accident
		first class of women at Yale	death of daughter
		pro football injury	being raped
		child's failure at school	winning a lottery
		teenage unpopularity	embezzlement
			first black women lawyer in the South
			blacklisted in Hollywood in the 40's

Table 3.2: Exemplary typology of life events based on the prototypical amount of exp[eriencers], prob[ability] of occurrence, and correlation with age (slightly adapted from Brim (1980, 153) by transposing the table and shortening the header)

(3.11) *[**work**.] All events that relate to happenings in one's main and secondary occupation. Includes both new professional activities, job loss, changes in the context of one's own career.*

(3.12) *[**education**.] All events that serve to prepare a career (training, further education, studies, degrees).*

(3.13) *[**family**.] (Event refers mainly to the degree of kinship) All events that are directly related to a related person: Birth, death, marriage of relatives, separation of relatives, family reunions, conflicts or reconciliations within the family.*

(3.14) *[**romantic**.] (Event refers mainly to the partnership with a person and not to the family relationship) All events that are directly related to partnership developments: Own wedding, engagement or marriage proposal, separation, new partner:in, [sic!] other relationship events.*

(3.15) *[**other_social**.] All social events that are not directly related to one's own family or partnership (friends, neighbours, acquaintances).*

(3.16) *[**leisure**.] All events related to personal leisure activities: Holidays, leisure activities (swimming, hiking...) Purchases for leisure activities, relaxation and recreation, hobbies*

(3.17) *[**money**.] All events primarily related to financial matters: personal insolvency, inheritance, debts, borrowing, taxation*

(3.18) *[**health**.] All events related to one's own health: Injuries, illness, recovery, hospitalisation, admission and termination of long-term treatment, diagnoses.*

(3.19) *[**housing and housing situation**.] All events related to the personal housing situation: moving, loss of the flat or house, alterations, renovations, flat hunting.*

(3.20) *[**Collective events**.] All events related to collective life events: war, pandemic, climate crisis, etc.*

(3.21) *[**Other domain**.] Code this when there is no clear way of coding the event, and building a new category is overly difficult*

(3.22) *[**Not clear: Poor information**.] Code this when the information provided for the event does[']nt allow a concise categorization of the event*

(3.23) *[**Not clear: more than one domain**.] Code this when information on more than one event is provided*

Whereas most finer-grained domains can be inferred from these definitions, it should be added that the class `Other domain` encompasses the subdomains **Religious/Spiritual event** as well as **Legal troubles** defined as follows:

(3.24) *[**Religious/Spiritual event**.] Code this event when an individual participates in a significant religious or spiritual event or ceremony. This can include religious rituals, ceremonies, or milestones in their faith or belief system.*

(3.25) *[**Legal troubles**.] Code this event when an individual encounters legal issues or challenges, such as involvement in legal disputes, facing legal consequences, or engaging with legal processes.*

The three classes `family`, `romantic`, and `other_social` are all seen as "Social" events and therefore share most of their subdomains. The main difference is the distance between the individual who experienced or took part in an event and the subject of the autobiographic report. This is hinted at by the respective clarifications in brackets in the Definitions (3.13), (3.14), and (3.15). With respect to the aforementioned event types, it is worth pointing out that most of the events mentioned in these definitions are actual *changes of state* of the subject, However, especially the *housing* domain may involve *process events*, too.

3.3.2 Computational methods and resources

Since events can be seen as one of the main building blocks of any narrative text, they play an integral role in the field of computational literary studies. In computational biographical research, by contrast, they play a minor role albeit that there are notable exceptions. Rehm et al. (2017), for instance, tried to extract movement-related events from early 20th century letters in order to automatically generate travelogues. Similarly, Stranisci et al. (2022) annotated biographical Wikipedia entries with different event types. As a matter of fact, these two research endeavors are prototypical for computational methods and resources dealing with events in the sense that the former involves an **event detection** and the latter an **event classification**. System-wise, most work is either based on various "traditional" machine learning or neural approaches.

Event (trigger) detection

Due to the aforementioned focus on verbs in event models from computational linguistics, it is not surprising that many studies try to identify **event triggers** based on this part of speech. In addition, some event trigger detection systems were also trained to annotate nouns (Upadhyay et al., 2016; Huang et al., 2018; Stranisci et al., 2022), or nouns and adjectives (Sims et al., 2019).

By contrast, the more generic **event detection** task is usually less restricted with respect to linguistic properties of the spans to be classified as being events (Araki et al., 2018; Caselli and Üstün, 2019). Regarding the typology of generic classification tasks (see Section 2.3), an event detection is typically rather operationalized as a token classification (Rehm et al., 2017) and not a text classification like the tasks of this thesis. Nevertheless, it should also be noted that many such studies combine the token classification (= extraction of relevant phrases) with a subsequent text classification (Jung and Stent, 2013; Caselli and Üstün, 2019; Vauth et al., 2021).

Event classification

Whereas event domains from biographical research like the one proposed by Haehner et al. (2024) have, to the best of my knowledge, not yet been implemented computationally as a text classification task, similar ones have. One example worth mentioning is the Automatic Content Extraction (ACE) standard which comprises the eight event types `LIFE`, `MOVEMENT`, `TRANSACTION`, `BUSINESS`, `CONFLICT`, `CONTACT`, `PERSONELL`, and `JUSTICE` with 33 subtypes in total.[6] As inferable from the labels, this classification task has primarily a news domain scope albeit that the subtypes `BE-BORN`, `MARRY`, `DIVORCE`, `INJURE`, and `DIE` of the first ACE event type are applicable to biographical sources. In addition, the ACE standard involves further tasks, such as the recognition of entities and relations (Doddington et al., 2004, 837). This is why models that combine them may solve complex tasks similar to an aspect level sentiment analysis. However, this kind of work is sidelined henceforth.

In computational literary studies, narrative classifications are more prominent than topical ones. Hence, the operationalization of the four event types `change_of_state`, `process_event`, `stative_event`, and `non_event` (Vauth et al., 2021; Gius and Vauth, 2022) mentioned earlier is somewhat prototypical for this field of research. Interestingly, Vauth et al. (2021, 337) assigned what they call "narrative values" (= 7, 5, 2, and 0) to these classes in order to assess the respective narrativity trajectory over the course of a given text. Figure 3.9 shows one such example which is in fact similar to a satisfaction chart used in biographical research (see Section 2.1.2) as well as sentiment trajectories (see Section 3.1.3).

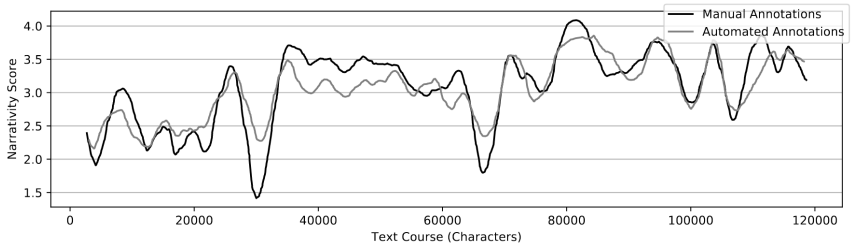

Figure 3.9: Manually and automatically annotated narrativity trajectory of Franz Kafka's *Die Verwandlung* (Vauth et al., 2021, 341)

Event classification systems

In their survey, Xiang and Wang (2019, 173130) compared 37 models with regard to their performance on distinguishing between the ACE event types and showed that neural approaches performed best with F1 scores in the range .69–.76 (mean .72), followed by a pattern-based system that yielded an F1 score of .70 and more "traditional" machine learning models with F1 scores in the range .60–.70 (mean .66). Vauth et al. (2021, 338–341) reached an weighted F1 score of .78 for their quaternary narrative event type classification with an transformer-based classifier. Caselli and Üstün (2019) used the seven event classes `OCCURRENCE`, `ASPECTUAL`, `PERCEPTION`, `REPORTING`, `INTENSIONAL_STATE`, `INTENSIONAL_ACTION`, and `STATE` and yielded F1 scores in the range .43–.74 depending on the amount of training data and language when fine-tuning the basic multilingual BERT model. They observed that an addition of Italian samples also improved their model fine-tuned for English texts while the opposite setting yielded only minimal gains. More specifically, the fully Italian version trained on 17,528 samples outperformed all off-the-shelf systems. However, 10,603 additional English fine-tuning instances resulted in a model that underperformed by a 3 % point margin compared to the English state-of-the-art. The latter was achieved by a more "traditional" machine learning system that relies on high level lexical semantic features (Caselli and Morante, 2018). Italian and Spanish models were also investigated by Sprugnoli and Tonelli (2017, 9–10) in order to evaluate their potential application on early 20th century Italian

[6] The most recent version is accessible via `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf` (Last accessed on February 18, 2026)

and 18th century Spanish datasets. As the latter time frame is similar to the one of this thesis, it is worth highlighting that they observed a notable F1 score drop to .39 which they attributed to "many diachronic language variations affecting precision".

In summary, current event classification systems yield lower performances than sentiment classifiers (see Section 3.1.2) – maybe because the former task involves more labels. Yet, the general tendency of neural approaches outperforming "traditional" machine learning ones stays the same. In a similar vein, both tasks have mainly been applied to contemporary sources albeit that more studies have applied a sentiment classification to historical ones than an event classification.

3.3.3 Typical issues and methodological critique

Typical operationalizations of event-related tasks tend to contrast potential application scenarios in historical research according to Sprugnoli and Tonelli (2017). To give an example, Rehm et al. (2017, 45–46) used an existing ACE dataset comprising modern-day English news articles to train and evaluate their system because of the scarcity of annotated in-domain data (= early 20th century letters). However, as the vast majority of their sources are German and a few French, their methodological approach required an automatic translation to English in many cases which is error-prone in itself. Additionally, they "discovered that several events could not be detected due to data formatting issues". This observation fits the critique of Sprugnoli and Tonelli (2017, 5) that "the complexity of [the] ACE annotation [scheme] makes the creation of consistent labeled data very challenging." Moreover, off-the-shelf models for event extraction are highly domain-specific and hence often not apt for an application on out-of-domain data (Ranade et al., 2022, 101587). This is especially true for an analysis of biographical sources since, for instance, the ACE standard includes a few relevant categories but still lacks some others. One example for the latter is the differentiation between the subject of a biography and other people mentioned.

Considering narrative event classifications, Vauth et al. (2021, 339) encountered the issue that they had too few training instances for one class which may be the reason why respective test instances were commonly attributed the majority class instead. However, this was not as problematic in this specific case because their downstream task relies on narrative values which were similar between most of the confused classes (Vauth et al., 2021, 340). Thus, there are fewer differences observable between their manual and automatic annotations than one might expect (see Figure 3.9). Still, Gius and Vauth (2022, 9) raised the question whether the exact "values are appropriate" although they see "an obvious ranking of categories."

Target-wise, the work described by Vauth et al. (2021) and Gius and Vauth (2022) aims at modeling plot and/or genre differences on the base of an eventfulness so that issues similar to an approximation based on sentiment polarity arise (see Section 3.1.4). At first glance, the similarity between the two approaches may be less apparent as an event-based one has mainly been applied to literary sources so far. However, the difference between fiction and non-fiction is somewhat blurred in the case of (historical) biographical sources (Nünning, 2022, 30). This is why the statement that it "is questionable whether the distinctive nature of a genre can be delineated so clearly from that of other genres or be captured in simple, general formulas of this kind" (Hühn, 2014, 168) also applies to these kinds of sources.

Finally, it is noteworthy that the heterogeneity of underlying concepts summarized under the term *event* is also sometimes criticized (Gius and Vauth 2022, 2; Ranade et al. 2022, 101580) which increases the similarity to sentiment analyses further. Hence, a more narrative approximation towards concepts, such as plot and genre, may be worth the effort but, when combined with sentiment polarity, does not add as much of a perspective to biographical research as a more topic-related categorization of life event domains. In a similar vein, a survey among history scholars showed that an event type classification is the most important event property in this application domain but only if it is based on semantics and not grammatical and/or syntactical features (Sprugnoli and Tonelli, 2017, 14–15).

Chapter 4

Biographical data under investigation

The biographical texts analyzed within this work are derived from two quite distinct data sources: On the one hand, there are 18th century memoirs of members of the Moravian Church and, on the other, entries in a biographical dictionary mostly compiled in the late 19th century. This work focuses primarily on the former. However, the latter are included in order to (i) highlight further application scenarios and (ii) stress typical differences between different types of historical biographical texts. Both data sources are described in the following subsections first in isolation and then in a comparative manner. More specifically, Section 4.1 deals with relevant information about Moravians and their memoirs whereas Section 4.2 lists characteristics of 19th century biographical dictionaries. Finally, Section 4.3 summarizes similarities and differences between the two biographical data sources before the subsequent Chapter 5 describes the compilation of a subcorpus that is used to train and evaluate classifiers as outlined in later chapters.

4.1 Moravians and their memoirs

The Moravian Church is a Christian denomination which was founded in the early 18th century and is also known as *Unitas Fratrum* 'Unity of Brethren' referencing an eponymous 15th century precursor. In the 18th century, lives of Moravians were influenced by certain circumstances some of which are highlighted in Section 4.1.1. These circumstances are also reflected in semi-autobiographical records called *memoirs* which differ systematically from similar sources (see Section 4.1.2). Moreover, the custom of writing memoirs has led to a comparably high amount of ego-documents that deal with Moravians – especially in the 18th century. This is why Section 4.1.3 describes how this amount is subsampled to the one used in later parts of this thesis.

4.1.1 Moravians in the 18th century

Precursor movements of the Moravian Church predate the Lutheran Reformation as they arose after the Bohemian reformer Hus had been executed in 1415 according to Atwood (2004, 21–22). Hence, they are among the oldest Protestant denominations which by extension also applies to the Moravian Church. This "Renewed Church" was founded by the Pietist Count Zinzendorf in the 1720s in Saxony by accepting religious refugees from Bohemia and Moravia. It is, however, not undisputed whether they "were indeed members of the old Unitas Fratrum" (Atwood, 2004, 21).[7]

The founding act illustrates that the inclusion of "all social strata and many nationalities", as Mettele (2007, 11) put it, has been some kind of defining feature of Moravians from the early 18th century onwards. Some diversity of national backgrounds arose from the religious refugees that were part of the first members of the denomination. Yet, even more arose from thousands of Moravians who migrated to North America and especially Pennsylvania with its constitutionally protected freedom of religion at that time (Atwood, 2004, 35–36).

[7] Interestingly, the English name *Moravian* focuses on the geographical origin of the refugees whereas its German counterpart *Herrnhuter* references one of the earliest settlements established by Zinzendorf.

Since Moravians conducted missionary work, they built settlements not only in continental Europe and the British Isles but also in Greenland and North America (Faull, 1992, 26). The same is true for Barbados in the Caribbean, Suriname in South America, Russia, and Tibet (Mettele, 2007, 13). However, missionary work involving Native Americans was reduced in 1788 after the founding of the United States of America due to legal constraints (Becker-Cantarino, 2016, 178).

Many Moravian settlements around the world were established and run on the base of so called *choirs*. On the one hand, this term refers to common buildings in which Moravians lived and, on the other, to social groups of individuals sharing the same gender and marital status as well as a similar age range (Mettele, 2007, 12). In some communities, the choir system followed communal principles and a strict segregation in the 18th century. One such example is Bethlehem, Pennsylvania, where Moravians lacked private property, and married couples were supposed to live separated from their partners and children between 1740 and 1762 (Faull, 1992, 26–27). During this time frame, young children were raised in nurseries although the separation from parents was subsequently postponed to the age of six and abandoned in the 19th century (Mettele, 2007, 12). With regard to family and romantic relationships, it is worth adding that spouses were typically drawn by lot in the 18th century (Becker-Cantarino, 2016, 185).

The lives of Moravians are said to have been less determined by gender than in other 18th century communities. This is due to the fact that women had similar occupations to men as pointed out by Faull (1992, 24) and Mettele (2007, 34). In a similar vein, social backgrounds are also deemed less important as, for instance, the Moravian church also "allowed peasants and artisans to hold office" from the very beginning (Atwood, 2004, 25). This custom may relate to the peculiarity that only a few members of this denomination had an academic background at that time.

As a result, Moravians experienced many housing changes on a local scale due to the choir system as well as on a global scale due to migratory movements and their missionary work. This is why it is not surprising that maps (Sciuchetti Jr., 2022) and travel diaries (Roseen, 1994) are among the many sources handed down from said time. In addition, the comparably high degree of equality between nationalities, gender, and social classes makes this denomination an interesting subject of inquiry for various fields of research. As these peculiarities are likely incorporated in one way or another in ego-documents, such as the Moravian memoirs, they are sources worth investigating, and the focus of the remainder of this section.

4.1.2 Distinct features of Moravian memoirs

> "Moravians had been encouraged since the 1740s to write or dictate their spiritual autobiographies, or Memorien, [sic!] which were completed after their deaths by a member of their Choir. These memoirs were read out at the believer's funeral and copied into the diaries of the respective Choirs."
>
> Van Gent (2017, 245)

This quote illustrates (i) the reason for writing, (ii) the writing process as well as the people involved in it, (iii) the expected audience, and finally (iv) the dissemination. These four aspects are elaborated further in the following:

From a theological standpoint, the **reason for writing** (and also collecting) memoirs emerged from the pietistic belief that exemplary life narratives are of value – no matter the social class or gender of an individual (Mettele, 2007, 15–16). Hence, this form of life writing can be seen as a continuation of a biographical tradition tied to funeral sermons (Schmid, 2004, 52). The same is true for various other types of ego-documents that emerged in the 17th century (Albrecht, 2022, 342).

In practice, Count von Zinzenberg was aware of these tendencies and is said to have initiated the Moravian custom to read out a retrospective life account during funeral services. The funeral setting led Böß (2016, 68) to count the practice of writing memoirs as a rite of passage between life and death. It also relates these sources to obituaries (see Section 2.1.1). Target-wise, this fits the claim that Moravian memoirs grant first and foremost each individual a way to "say goodbye", and,

secondly, serve historiographical purposes for the denomination as a whole (Faull, 1992, 32). These objectives explain why a memoir is usually written in old age "when its author has the opportunity to look back on life" (Faull, 1992, 32). In this context, it is worth highlighting that women often started writing their memoir "before marriage" (Faull, 1992, 32) which might be related to a high maternal mortality rate and, hence, fitting the aforementioned main reason for writing.

As for the **writing process**, Becker-Cantarino (2016, 179) noted that especially the first memoirs from the 1750s consisted mainly of protocol-like notes of the main biographical dates and events. Later memoirs included an increasing amount of more personal (auto)biographical content. Still, the final text was always composed by an unknown choir helper – an administrative position in the Moravian church – on the base of (auto)biographical drafts. This person and/or family members added an account on the death circumstances and sometimes also a preface with basic demographic information (Faull, 2021, 220–221). This is why an authorship attribution is rather difficult. The same is true for the categorization of these sources as being biographical and/or autobiographical which is also the case for many other pietistic texts with (auto)biographical content from that specific time frame (Albrecht, 2022, 342).

Considering the **target audience**, it is likely that most (co-)authors were familiar with the custom of reading out the final version as part of the funeral service. This is why it is not surprising that many memoirs as well as preparatory notes include Bible verses thematically fitting said occasion (Böß, 2016, 75–76). Similarly, it can be seen as a known fact that the only "taboo was to provide a narrative that did not lean towards salvation but rather to estrangement from the community" (Mettele, 2007, 23). However, such specifications were rather implicit and may have arisen from the perception of other memoirs since Becker-Cantarino (2016, 188) claimed that no explicit instructions on how to write these ego-documents have been handed down. This fact could relate to the importance of individuality in the belief system of Moravians which has roots in the Pietist tradition. Still, Böß (2016, 69) pointed out that at least from 1822 onwards, pupils have practiced writing memoirs in Moravian schools. It is, however, unclear how much these exercises impacted the autobiographical draft typically written much later in life.

With regard to the **dissemination** and archival of memoirs it is worth highlighting that each text exists potentially in multiple versions, and that they may differ from each other significantly: The first version is the original autobiographical draft or preparatory notes written by the subject albeit that some texts were written by a different person from the beginning. One possible reason for the latter is that some Moravians did not deem themselves proficient enough (Böß, 2016, 77). It should be noted that memoirs of the latter kind typically still follow a first person narration whereas a third person perspective was preferred "when an individual died without leaving a memoir" (McGuire, 2021, 33). Nevertheless, there are deviations from this general tendency. Faull (1992, 36), for instance, identified an overall diachronic stylistic change from a first to a third person narration. Thus, although an authorship attribution based on pronoun usage may seem plausible, it is not as easily applicable to these sources. No matter the text genesis, choir helpers usually conflated the initial (auto)biographical notes and added a final section about death in order to finalize the version read out at the funeral.

In some cases, a second handwritten version was derived from the original one and published in the so called 'Diary' which "recorded all comings and goings, religious services, births, deaths, and marriages" (Faull, 1992, 34).

Selected memoirs also exist in a third version which was shortened even further and distributed globally via the monthly periodical *Gemeinnachrichten* 'Congregational Reports' (Mettele, 2007, 17). This version stayed handwritten until 1818 when a print publication became the standard. Still, each issue comprised only two to three memoirs so that selection criteria were necessary and involved the recency of death, the importance of the person, and "a certain entertainment value" (Mettele, 2007, 19).

It needs to be added that some memoirs may have been censored or even destroyed after being archived. According to Peucker (2012, 181–182) this is due to the fact that Moravians tried to "protect the privacy of individual members of the church" and to not "depict Zinzendorf in an unfavorable light" after the General Synod in 1764. In this context, they wanted to reduce the

amount of sources relatable to "elements of radical Pietism or enthusiasm" (Peucker, 2012, 195), too. It is, however, unclear to what extent this affected the various versions of Moravian memoirs.

In summary, this complex text genesis leads to a situation where an authorship attribution is difficult if not impossible to accomplish. The same is true for the typological categorization of these sources as either biographies or autobiographies (see Section 2.1.1) since Moravian memoirs are better called "semi-autobiographical" (Brookshire and Reiter, 2024, 91). The subgenre status led Becker-Cantarino (2016, 174) to refer to them with the term *Aufsatz* 'written Account', and others with *Lebenslauf* 'life course' (Faull, 1992; McGuire, 2021) as Moravian memoirs usually start with a phrase comprising one of these two terms.

Language usage and stylistic features

Since many Moravians had a German-speaking background in the 17th and early 18th century, most memoirs were written in German until 1822 (Becker-Cantarino, 2016, 180). Yet, substrates from other languages as well as regio- and sociolects are likely to occur because of the various geographic and social backgrounds (see Section 4.1.1). Such linguistic peculiarities may have led Becker-Cantarino (2016, 183) to state that early memoirs are characterized by what she called *ungelenke Sprache* 'clumsy language'. This changes in memoirs from later decades – probably because of the aforementioned writing lessons in Moravian schools as well as the increased likelyhood of (co-)authors having read other memoirs.

Still, these texts comprise many instances of domain-specific language which led Peucker (2023) to even compile a Moravian dictionary. The additional senses of *choir* mentioned earlier are one such example but there are also typical abbreviations like *Hld* for *Heiland* 'savior'. Furthermore, Roth (2025, 147–159) observed a notable amount of hapax legomena due to very productive word formation patterns, such as ones based on *Herz* 'heart'. In addition, medical terminology is quite distinct in these sources as, for instance, *asthma* was called *Engigkeit* 'narrowness' while the term for non-specific health issues was *Zufälle* 'coincidences' (Roth, 2025, 189–195). Further peculiarities like the high degree of graphematic variation and a "very inconsistent use of punctuation" (Faull and McGuire, 2022, 136) arise from the respective time frame.

One typical **stylistic feature** is that Moravian memoirs are written in a highly emotional way (Van Gent, 2017, 245) but also in the style of a factual report with many dates (Roth, 2025, 202). This slight contradiction coincides with the field of tension between describing "personal detail interwoven with secular experiences" while also "repeat[ing] to an extent certain phrases, usually from the Bible" (Faull, 1992, 31). Hence, it is possible that part of the emotionality arises from the inclusion of religious verses which are emotional on their own. This is especially true for references to Christ's suffering. These have become somewhat obligatory at a certain point in time that they were apparently added at later stages of the writing process if missing (Böß, 2016, 222). In this regard, Faull (1992, 36–37) noted that an "ultra-realistic depiction of Christ's wounds" called "blood-and-wounds vocabulary" was common in Moravian sources in general although being officially forbidden after 1751.

Similarly, death descriptions are characterized by many positively connoted euphemisms like *Vollendung* 'completion' or *seliges Ende* 'blessed end' (Roth, 2025, 195). By contrast, common phrases like *seliger Bruder* 'blessed brother' may seem emotional or religious at first glance. However, they are used so phrasally in these sources that they are best deemed as following a rather neutral tone (Becker-Cantarino, 2016, 182). Finally, it should be added that Moravian memoirs contain a mixture of content elements "from professional biographies, travel journals, adventure stories, or even captivity narratives" (Mettele, 2007, 29) with no apparent gender-specific differences (Mettele, 2007, 33–34).

Basic structure

In general, Moravian memoirs "depict the spiritual growth of the individual" with an "interweaving of spiritual introspection and secular experience" in the words of Faull (1992, 43). Due to the pietistic tradition of valuing individual differences, they vary in length. Shorter ones span two

handwritten pages whereas longer ones exceed 30 (Schmid, 2004, 51). The average length was 15 pages in the octavo format in the 18th century (Mettele, 2007, 14). Still, all memoirs can be expected to portray the most important life events chronologically which is why they mostly follow the same basic structure comprising the six sections summarized in Table 4.1:

Section	**Expectable life events** (Becker-Cantarino, 2016, 181–182)	**Expected polarity** (McGuire, 2021, 133)
Demographic facts	birth, baptism	"neutral"
Childhood	education	"positive peak"
Adolescence	various challenges	"fall"
Conversion experience	conversion, migration	"rise slightly"
Midlife	occupational changes, marriage, child birth	"remain positive"
"Going home"	illnesses, death	"final peak"

Table 4.1: Overview over expectable life events and their polarity per typical section of Moravian memoirs with references to life stages in italics

The first section deals with basic **demographic facts**, such as the names and confessional background of the parents as well as the birth and baptism of the biographical subject (Becker-Cantarino, 2016, 181). After these introductionary notes, **childhood** is portrayed with a focus on "innocence" as Faull (1992, 32) put it. This second section may also include some notes on family upbringing and school education according to Becker-Cantarino (2016, 181). Moreover, McGuire (2021, 133) suggested that there should be a "positive peak in early childhood" that "should then fall through adolescence and young adulthood [...] as people describe challenging experiences."

It may have been intentional that the third section that deals with **adolescence** is one Faull (1992, 32) called "troubled" as it contrasts the subsequently described **conversion experience** polarity-wise. The latter can be seen as one important biographical turning point although it was typically not as wordly described (Becker-Cantarino, 2016, 181–182). However, Böß (2016, 56) pointed out that an actual conversion could take many years so that this section may be described longer and could include some negativity in such cases, as well. The respective fourth section is also notable because a conversion could have happened at any age in theory so that it is not as easily relatable to an age group. Furthermore, some memoirs are about individuals who had been Moravians by birth due to the confession of their parents. It is, however, unclear how this influenced said section which is seen as obligatory in related work. In a similar vein, some memoirs deal with migration experiences while others do not – simply because the respective individual stayed in the same area for her or his entire lifetime.

McGuire (2021, 133) expected a description of **midlife** characterized by "stability" as a fifth section. It typically follows the ups and downs of the preceding ones, and is still portrayed positively but to a lesser degree. This section is likely to include references to occupational and housing changes. The latter are mostly related to the missionary work in the 18th century (see Section 4.1.1). In addition, this part of a memoir deals with major family events like marriage and child birth, if they occurred. However, it should be noted again that, at that time, future spouses were typically drawn by lot and lived separated from each other and possible children. This is why said events had fewer effects on the daily life than in other time periods and communities.

Finally, all memoirs end with a **"going home"** section (McGuire, 2021, 133) that focuses on final illnesses and the eventual death. Both events are typically depicted as a positively connoted "trial" one needs to overcome. The importance of this sixth section is underlined by the fact that it is sometimes longer than the rest of the text according to Schmid (2004, 51) and Böß (2016, 238).

4.1.3 Identification of relevant memoirs

Due to the importance of memoirs as a rite of passage between life and death, the amount of texts is comparably high for 18th century sources. For instance, the Unity Archives in Herrnhut list 29,063

memoirs in their online catalogue,[8] and the Moravian Archives, Bethlehem, 6847.[9] In the case of the German archive, 26,982 (= 92.8 %) are about people born and/or deceased between 1750 and 1850. The same is true for 5822 (= 85.0 %) memoirs from Bethlehem. However, it is unclear how many of these are German texts because the inventory lists only include basic descriptive metadata like the subject's name as well as dates of birth and death if known.

Some memoirs have been published as textual surrogates – mostly in the form of scholarly editions – over the past decades. One example is the work of Lost (2007) which comprises transcriptions of 31 German memoirs from the 18th and 19th century in print. Another print edition is the translation of 30 German memoirs from the Moravian archives in Bethlehem, Pennsylvania, into English by Faull (1997). However, Faull (2021, 216) estimated that all these printed surrogates account for no more than 10 % of the memoirs composed between 1750 and 1850 in total so that most memoirs remain in archives.

Considering digital editions, one of the first ones was the *Bethlehem Digital History Project (BDHP)* which started in 1999 and currently offers digital access to transcriptions of 34 memoirs.[10] Some of the memoirs are German ones and enriched with English translations while others have been English from the beginning. The most ambitious ongoing digitization project is called *Moravian Lives*[11] and conducted collaboratively by the Centre for Critical Heritage Studies at the University of Gothenburg, Sweden, and Bucknell University in the United States since 2015. It grants access to metadata about 60,000 memoirs from various Moravian archives. In addition, it involves a crowdsourcing project in which a steadily increasing subcorpus is digitally transcribed. To some extent, both digital edition projects focus on the early days of the Moravian missionary work in North America. This is why they deal with ego-documents of many German speaking immigrants (see 4.1.1). Both this time frame and geographic area reflect interdisciplinary research interests because they coincide not only with the early years of Moravian missionary work but also collective events like the American Revolutionary War (1775–1783) and the subsequent formation of the United States. This is also the reason why these digital editions were used as a starting point for the selection of relevant memoirs here, and amplified by further (related) projects. The main sampling criterion was that the biographical subject has lived in the 18th century. To ensure this, only sources of people born and/or deceased between 1750 and 1850 were considered. This led to a sample of digital transcriptions from the following projects:

BDHP. The 17 German biographies from the Bethlehem archives found on the BDHP website which are not also part of the *moravian-lives_de* dataset in order to avoid duplicates.

bethlehemer-schwestern. 35 German memoirs of women from the Bethlehem archives that are part of Faull's project *Umgang mit dem Heiland: Lebensläufe der Herrnhuter Schwestern aus Nordamerika* 'Dealings with the Savior: Memoirs of North American Moravian Sisters'.[12] 30 of these have previously been translated into English by Faull (1997).

moravian-mens-memoirs. 5 German memoirs from the Bethlehem archives that are part of Faull's project *Moravian Men's Memoirs*.[13] According to the website, all of them have also been published in print.

moravian-lives_de. 32 German memoirs from the Bethlehem archives that have been made accessible via the *Moravian Lives* website[14] over the past years. 300 more were marked as being in the process of digitization when the respective note as well as all the memoirs selected here were findable before the website reboot at the end of 2024.

In total, this corpus of 89 memoirs comprises 137,092 tokens or 20,305 types (see Table 4.2). For each memoir the subject's name as well as year and place of birth and death are available as metadata – partly due to further research conducted while sampling process. For instance, the place

[8] https://www.unitaetsarchiv.findbuch.net/ (Last accessed on February 18, 2026)
[9] https://www.moravianchurcharchives.findbuch.net/ (Last accessed on February 18, 2026)
[10] http://bdhp.moravian.edu/personal_papers/memoirs/memoirs.html (Last accessed on February 18, 2026)
[11] http://moravianlives.org/ (Last accessed on February 18, 2026)
[12] https://katiefaull.com/moravian-materials/ (Last accessed on February 18, 2026)
[13] https://katiefaull.com/moravian-materials/ (Last accessed on February 18, 2026)
[14] http://moravianlives.org/ (Last accessed on February 18, 2026)

of birth was not available off-the-shelf in all cases but inferable from the first one or two sentences in a close reading step because of the basic structure (see Table 4.1).

Dataset	Documents	Tokens	Types
BDHP	17	37,513	8,838
bethlehemer-schwestern	35	45,325	9,008
moravian-mens-memoirs	5	16,587	4,545
moravian-lives_de	32	37,667	8,081
Total	89	137,092	20,305

Table 4.2: Basic corpus statistics of Moravian datasets

4.2 Biographical dictionary entries

The growing biographism in the 18th and 19th century was not restricted to the religious domain as it also led to the emergence of biographical dictionaries which were typically rooted in a growing nationalism, too. One of the first large scale endeavors of this kind was the so called *Allgemeine Deutsche Biographie (ADB)* 'General German Biography'. It was compiled by the Historical Commission of the Bavarian Academy of Sciences between 1875 and 1912. As a biographical reference work with a national focus, it shows typical features that will be described in Section 4.2.1 before Section 4.2.2 explains which entries will be analyzed within this thesis.

4.2.1 Distinct features of biographical dictionary entries

In the 19th century, large scale biographical dictionary projects aimed at the development of a national consciousness in many European states, such as Sweden, the Netherlands, Belgium, Germany, and Great Britain (Hockerts, 2008, 232). Of these, the Belgian national biographical dictionary is said to have influenced the German ADB the most which in turn influenced the British one. In the case of the ADB, the term *national* was defined broadly since it was not bound by political borders but rather used in the sense of a language area and cultural space. In addition, Hockerts (2008, 234) claimed that it incorporated founding myths of the German Empire in order to suggest a millennium of national continuity. Yet, nationalism was not the only **reason for compiling** the ADB as, for instance, plans to avoid the marginalization of Catholics and women were also discussed during the preparation phase of the project in the 1860s (Hockerts, 2008, 229–230). In this sense, there are parallels to the aforementioned fact that Moravians wrote memoirs in order to value lives of women as well as individuals of different social classes (see Section 4.1.2).

The ADB focused on a critical academic assessment of the contribution of "important individuals".[15] This academic ambition explains why biographies of people who became famous in fields like history, science, art, and economics predominate (Hockerts, 2008, 234). These fields are somewhat mirrored by the selection of **authors** who predominantly had an academic, artistic, or literary background, too (Hockerts, 2008, 235). Since most authors wrote biographies about people they knew personally and/or cherished, they sometimes wrote comparably long ones. A notable example is the entry about Bismarck which spans 225 pages (Hockerts, 2008, 237). This can be seen as a sign of discipleship which is a typical feature of biographical sources (Cockshut, 2001, 79). Furthermore, authors typically shared a socio-cultural background with the people they wrote biographies about. In this way, the ADB offered many different minorities and milieus, other than the socialist one, a chance for self-portrayal (Hockerts, 2008, 238). Both practices resulted in a total amount of 1850 authors who published 3548 entries in the ADB (Hockerts, 2008, 235). Although most entries were only written by one signing author, it is possible that further associates were involved in the preparation and/or writing process.

[15] The original German quote is "alle bedeutenderen Persönlichkeiten" (Hockerts, 2008, 234).

Target-wise, the academic assessment fits the **intended readership**. More specifically, the ADB was mainly compiled for further academic inquiry as well as the broader group of "the entirety of educated people".[16] All these characteristics led Reinert and Scholz (2017, 1) to state that ADB entries do "not only inform about [...] subject[s], but they also give insight into the authors, their questions, methods, intentions, biases, and the zeitgeist." Hence, these biographical texts may in fact become interesting sources for research dedicated to the history of science on their own (Reinert and Scholz, 2017, 1).

Dissemination and digitization

Similar to other projects of this kind, ADB entries were originally written in alphabetical order. They were published in 45 volumes between 1875 and 1899, followed by 10 further supplementary volumes until 1910, and, finally, an index of persons in 1912 (Hockerts, 2008, 235). Especially the supplementary volumes added revisions of previously published entries so that some entries exist in multiple versions. In addition, they included entries about people who died after the publication of the volume they would fit in alphabetically if they were deemed important enough. Thus, not all entries were actually published in a strict alphabetical order. This is even more true for later additions and revisions within the context of digitization projects: In 1999, facsimiles of all ADB entries became available online. The index of persons was added in 2001 with additional metadata, and machine-readable versions of the entries in 2010 (Reinert and Scholz, 2017, 4).

Textual characteristics

With regard to **linguistic particularities**, it can be expected that there are only a few deviations from the grammatical and graphematic standard of the German language at that time. This is due to the academic setting and the ambition to write "in a highly legible form".[17] However, this standard is still distinct from the modern one since Stotz et al. (2015, 75) noted that the biographies "are written in an outdated orthography and style".

Considering **stylistic features**, ADB entries follow a third person narration and depict the most important life events and achievements chronologically in order to follow the reference work nature. In a similar vein, a lot of entries start with basic demographic facts like places of birth and death that have become part of the headline in an ongoing successor project (Reinert et al., 2015, 16). Other structural elements, such as a genealogical abstract or a summary of the most important work and achievements (Reinert and Scholz, 2017, 5), may appear in some ADB entries while missing in others. This variability is due to the high degree of structural and content-related freedom ADB authors were granted so that the entries are, as a matter of fact, more akin to literary essays than objective life descriptions (Richter and Hamacher, 2022, 200).

4.2.2 Identification of relevant entries

In total, the ADB comprises 26,300 biographical entries about people born before the year 1900 all of which were originally distributed in printed form but subsequently digitized (Hockerts, 2008, 262). Therefore, all entries are currently also available in digital form. Within the digitization project, metadata fields like the subject's name and profession as well as the year of birth and death were extracted in order to be included in a faceted search (Reinert et al., 2015, 15). Said search capacity of the ADB website[18] was exploited here in order to subsample the initial corpus of all ADB entries to one which only deals with individuals that lived in the 18th century. More specifically, this target time frame was operationalized as a search for people whose year of birth and death both lie between 1750 and 1850 – a range of years similar to the one used for Moravian memoirs as described in Section 4.1.3.

[16] The original German quote is "die Gesammtheit [sic!] der Gebildeten" (Hockerts, 2008, 234).
[17] The original German quote is "in wohllesbarer [sic!] Form" (Hockerts, 2008, 234).
[18] https://www.deutsche-biographie.de/extendedsearch (Last accessed on February 18, 2026)

This one and only subsampling step decreases the number of biographies to 3548 (after the removal of duplicates) which comprise almost 3.1 million tokens or 255,418 types in total (see Table 4.3). The fact that the selected amount of documents equals only 13.5 % of the whole ADB underlines that most entries of this reference work are about people who lived in earlier (or in a few cases more recent) times. This fits the aforementioned aim of accounting for a whole millennium of national continuity and not only focusing on recent celebrities. The other aim of not marginalizing women was, however, not achieved since only 102 entries (= 4 %) from the selected time frame have female subjects.

Dataset	Documents	Tokens	Types
ADB	3,548	3,072,662	255,418

Table 4.3: Statistics of the ADB dataset

4.3 Similarities and differences

The information on Moravian memoirs and ADB entries discussed in Section 4.1 and 4.2 are summarized in Table 4.4. It is noteworthy that more 18th century Moravian memoirs have been handed down to this day than ADB entries although the exact number of German memoirs is difficult to assess. Yet, only a small fraction of memoirs are available in digital form whereas all ADB entries are accessible in this way. One reason for this significant disparity may be the fact that the former are mostly handwritten (when disregarding versions printed in journals) and, hence, more difficult to transcribe automatically than the latter which were typeset (see Section 2.1.2).

Both types of biographical sources share the possibility that multiple versions of the same text exist while an authorship attribution is only difficult in the case of Moravian memoirs. Considering the time of origin, all texts from both corpora were finalized posthumous. Yet, Moravian memoirs were written decades earlier and closer to the subject's death in order to be read out during the respective funeral. Such a religious reason for writing is not present in the case of ADB entries. Instead, they were written for the purpose of building a national consciousness on the one hand, and academic interests on the other. The latter is interrelated with the target audience which is a peer group of the authors. Interestingly, this situation resembles the case of Moravian memoirs, even though both groups differ from each other. More specifically, the ADB focus more on "important" people while Moravians wrote more about "common" ones. Hence, it is not surprising that all these differences and similarities are reflected in the language of these texts, too.

	Moravian memoirs	**ADB entries**
Sources handed down	32,802	3,548
Sources available	89	3,548
Form	mostly handwritten	typeset
Versions	up to three	revisions possible
Author(s)	a few (mostly including self)	typically one
Time of writing	mostly 18th century	1875–1912
Reason for writing	religious	nationalist/academic
Target audience	other Moravians	mostly academic
Scope	all Moravians	"important individuals"
Language	regio-/sociolectal	academic

Table 4.4: Comparative overview over distinct features of Moravian memoirs and ADB entries about people who lived between 1750 and 1850

It is possible to relate some of the aforementioned differences to common types of biographical texts and their distance between author and subject on the one hand as well as their totality ambition

on the other (see Section 2.1.1). As illustrated by Figure 4.1, Moravian memoirs are more similar to obituaries than to prototypical memoirs based on these features. Similarly, ADB entries are less distant than expectable of (modern-day) dictionary entries.

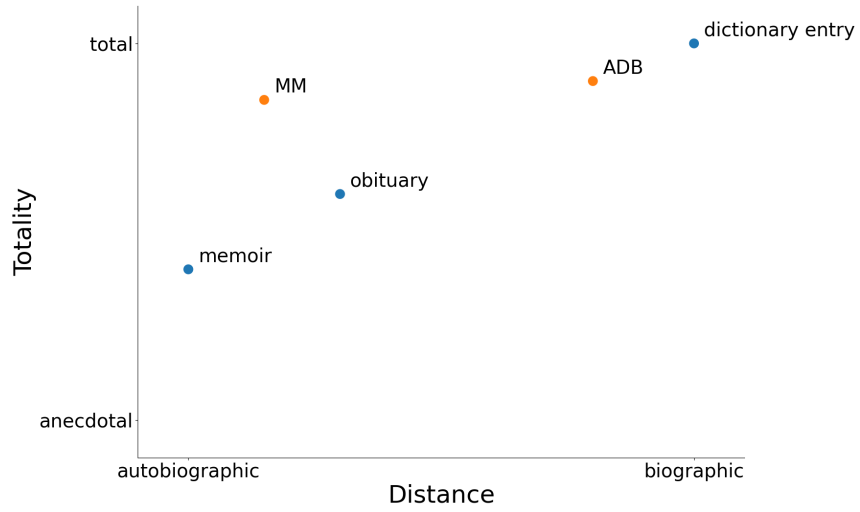

Figure 4.1: Idealized distance and totality of selected biographical subgenres (in blue) compared to Moravian memoirs and ADB entries (in orange)

Textual differences

In order to assess textual differences – and in extension genre as well as topic differences that may coincide with them –, text embeddings were investigated first. To this end, all 89 Moravian memoirs and 3548 ADB entries were vectorized with (i) TF-IDF when considering only the top 1000 most relevant word tokens in total and (ii) the transformer model gbert.

Afterwards, the 1000 TF-IDF dimensions and 768 of gbert were both reduced to two dimensions with PCA in order to enable a visual comparison (see Figure 4.2). The resulting distributions imply that the two datasets are quite different from each other when using a TF-IDF vectorization but more overlapping when using gbert. Given that transformer embeddings tend to encode semantic relationships and not only n-gram statistics, this methodology-related difference may hint at different overt expressions for similar covert contents. This potential finding can be related to textual (sub)domain differences.

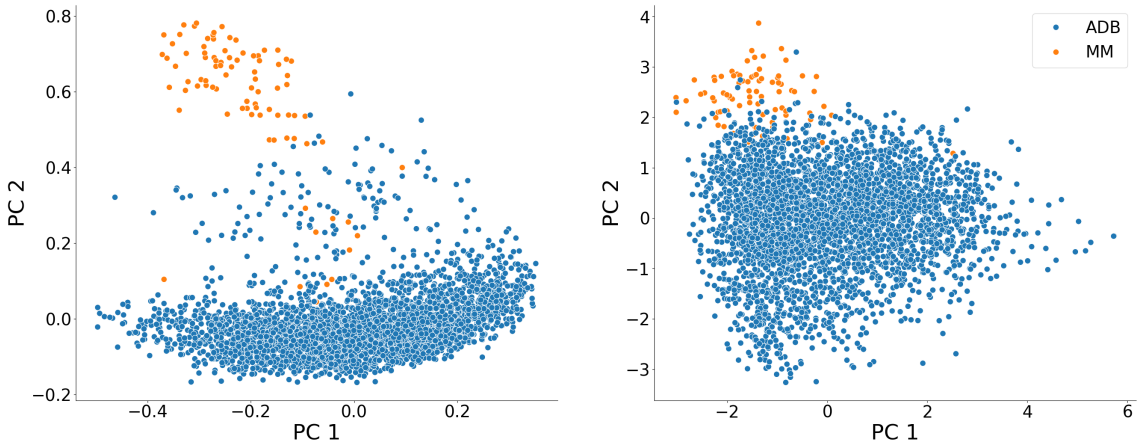

Figure 4.2: 89 Moravian memoirs and 3548 ADB entries when embedded with TF-IDF (left) and gbert (right) and projected to 2D with PCA

In order to further assess domain-specific language usage, TF-IDF was again applied. However, this second analysis followed its more prototypical application scenario of identifying relevant

tokens in order to distinguish between different documents or corpora. To this end, the lower cased word forms with the highest TF-IDF score per corpus were compiled and the top 30 are listed in Table 4.5.

Moravian memoirs	**ADB entries**
dabey, hld, wolte, konte, gieng, geschw, solte, lezten, kan, led, muste, ofte, laßen, herrnhaag, lezte, brr, herze, wobey, hhuth, july, krigte, america, juny, hlds, –, alleine, 6ten, loosung, litiz, gnadenhütten	k, universität, deutschen, wien, stellung, †, „die, namentlich, karl, sowie, französischen, werke, philosophie, „ueber, z, (s, zunächst, münchen, theologie, 1813, preußischen, titel, „der, göttingen, kunst, erschienen, studium, preußen, gymnasium, akademie

Table 4.5: Top 30 most relevant word forms per corpus

It is striking that the words that distinguish Moravian memoirs from ADB entries the most comprise many instances of graphematic variation like *dabey/dabei* 'thereby', *wol[l]te* 'wanted to', *kon[n]te* 'could', *gieng/ging* 'went', *sol[l]te* 'should', *le[t]zten* 'last', *kan[n]* 'can', *mus[s]te* 'had to', *ofte/oft* 'often', *laßen/lassen* 'let', *wobey/wobei* 'whereby', *July/Juli* 'July', *kri[e]gte* 'got', *America/Amerika* 'America', *Juny/Juni* 'June', and *Loosung/Losung* 'watchword'. Given that multilingual language model will play a role in later parts of this thesis, it should be noted in this context that some of these deviations from the modern German orthography make the respective word forms more akin to their English counterparts.

Furthermore, there are typical abreviations like *H[ei]l[an]d* 'savior', *Geschw[ister]* 'brethren' or 'sisters', *led[ig]* 'unmarried', *Br[üde]r* 'brethren', and *H[errn]huth* 'Herrnhut'. The latter one is a Moravian settlement just like *Herrnhaag*, *Liti[t]z*, and *Gnadenhütten* so that these two groups are somewhat expectable.

In a similar vein, the list of words that distinguish ADB entries from Moravian memoirs also comprises abbreviations like *k[aiserlich]/k[öniglich]* 'imperial/royal', *z[u]* 'to/for/near', and *s[iehe]* 'see', the metamark *†* 'died', and toponyms like *Wien* 'Vienna', *München* 'Munich', *Göttingen* 'Göttingen' and *Preußen* 'Prussia'. In addition, there are adjectives for nationalities like *deutschen* 'German', *französischen* 'French', and *preußischen* 'Prussian'.

Graphematic variation is only represented by the word *ueber/über* 'over' whereas the majority of the remaining word forms can be related to the academic domain like *Universität* 'university', *Werke* 'works', *Philosophie* 'philosophy', *Theologie* 'theology', *Kunst* 'art', *Studium* 'studies', *Gymnasium* 'high school', and *Akademie* 'academy'.

Since many of these findings fit the assumed geographic, linguistic, and social background differences surprisingly well, it becomes apparent that at least some of these more covert features are assessable on the base of textual surface forms. This has implications for the classification experiments and content analyses discussed in the following chapters.

Temporal differences

Besides textual features, temporal information exists for both corpora in the form of birth and death years. These are part of the metadata used to identify relevant ego-documents and said preselection required a slight difference between the two corpora: In order to keep as many Moravian memoirs in the corpus as possible despite their scarceness in digital form, sources were considered when they deal with individuals born *and/or* deceased between 1750 and 1850. However, ADB entries were only chosen when concerning people born *and* deceased between 1750 and 1850 in order to stay in the range of possibilities the faceted search of the ADB website offers. As illustrated by Figure 4.3, this led to an ADB sample which was not only written at a later time but also covers people who lived a little later. In addition, ADB entries consider fewer subjects who died in earlier life stages whereas the ages at death are more equally distributed in all 5833 Moravian memoirs collected in Bethlehem with metadata (see Figure 4.4). This applies to the 89 with fulltext digitizations as well, as they cover an age range of 4–95 years instead of 17–123 in the case of the ADB sample.

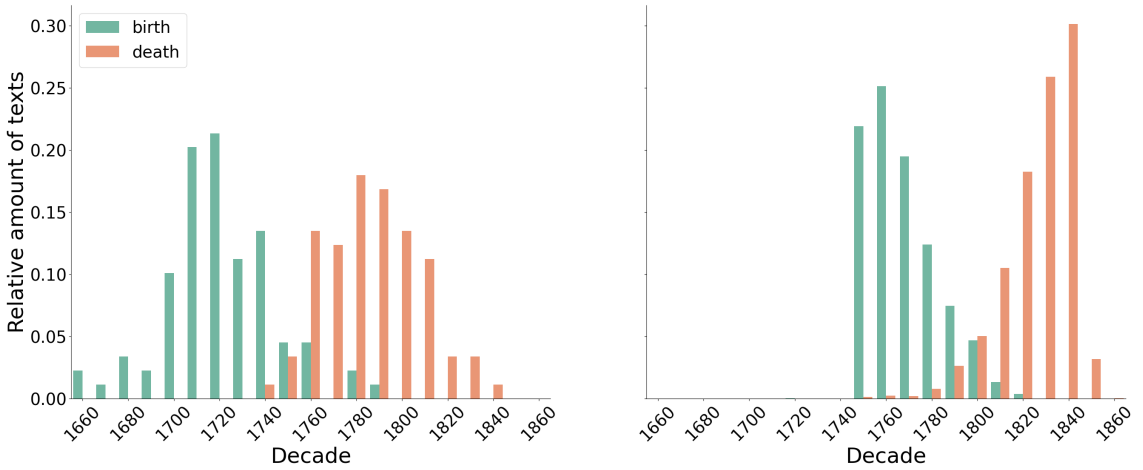

Figure 4.3: Relative number of births and deaths in 89 Moravian memoirs (left; with 2 unknown births and deaths) and 3548 ADB entries (right; with 256 unknown births and 107 unknowns deaths)

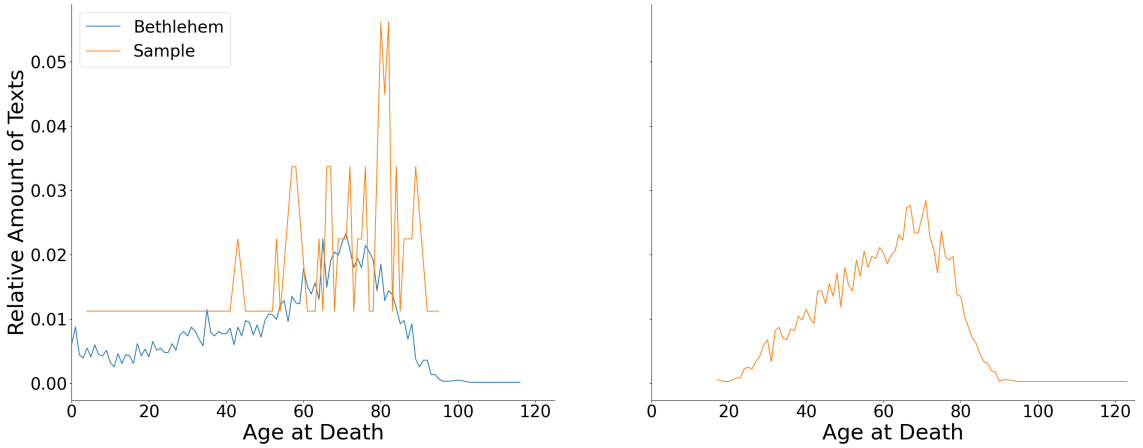

Figure 4.4: Age at death distribution in 89 selected Moravian memoirs (left in orange) out of 5822 from Bethlehem (left in blue; with 3/454 unknown ages) and 3548 ADB entries (right; with 300 unknown ages)

Gender differences

Whereas the gender of the authors is difficult to assess, the gender of the biographical subjects is given as metadata. As somewhat expectable from the respective occupational backgrounds, there is a male bias in the ADB corpus as only 4.0 % (= 102) of the 3548 ADB entries deal with women. By contrast, 60.1 % (= 54) of the 89 digitally available Moravian memoirs describe female lives so that there is a (less pronounced) female bias. This latter disparity is partly due to the lack of men's memoirs in the bethlehemer-schwestern subcorpus.

In conclusion, a focus on Moravian memoirs prevents the typical problem that biographical research tends to be skewed towards upper and middle class male people. However, the amount of German memoirs available in digital form is currently limited and they mainly deal with a very specific group of people from a very specific place. Therefore, it may also be useful to compare the depiction of their lives with the larger amount of ones from a biographical dictionary like the ADB in order to identify further differences and similarities some of which will be uncovered in Section 9.4.

Chapter 5

Corpus construction

A common dataset for all three text classification tasks described in the following Chapters 6, 7, and 8 was subsampled from the Moravian sources described in Section 4.1.3. The subsampling was conducted in a semi-random stratified fashion. To this end, available metadata fields were first enriched with geographical information so that the subsampling could account for this proxy of linguistic variation while also balancing gender distributions. Afterwards, the selected texts were tokenized into sentences and finally anonymized.

Metadata enrichment

Section 4.1.2 noted that each Moravian memoir from the time frame to be analyzed can be expected to follow a certain structure with basic biographic "metadata" in the very beginning. Said structure was exploited in order to identify birth places proxying sociolinguistic backgrounds. In practice, this means that a list of birth places was compiled by reading the first few sentences. As expected, the birth place was named in the beginning of the narrative in all cases but one. In the case of the sole exception, the whole memoir was read. However, it did in fact not include any information about the place of birth which is why it was excluded from the subsampling process. The birth place list was then geolocalized by reconciling it with the fuzzy search of the *geonames* database[19] so that it was possible to add latitude and longitude coordinates in 65 cases. The remaining 23 cases required more research. One reason was that some descriptions are rather generic, such as *in der Grafschaft Witgenstein* 'in the county of Witgenstein'. Furthermore, there were place names of settlements that no longer exist, such as *Heerendijk* (a former Moravian settlement in the Netherlands). Similarly, there were complex interdependencies between language change and graphematic variation like in the case of *Conne Walte* for *Kunewald* which is now called *Kunín*. Yet, it was possible to geolocalize all named birth places so that uncertainty remains only in the case of the five generic birth region descriptions where the actual birth place can only be approximated.

Subsampling procedure

The additional geographical metadata fields were considered when subsampling in order to ensure that as many dialectal areas as possible stay represented. This was achieved in most cases as illustrated by Figure 5.1. Said figure also shows that most memoirs are about people born in Europe and evidently mostly in communities that spoke some variety of the German language at that time. This is noteworthy because the texts were collected in Pennsylvania indicating that most if not all subjects died there. Still, there are a few people who may have learned German at a later stage of their lives. For example, one subject from the subsample (and two more from the full one) were born in England. Nine others from the subsample (and 29 from the full one) stayed their whole life in the New York/Pennsilvania area with some of them being Native Americans. Furthermore, the subsample comprises sources of two former slaves displaced from Africa, and the full sample another one about an individual born in the Caribic.

[19] https://www.geonames.org/ (Last accessed on February 18, 2026)

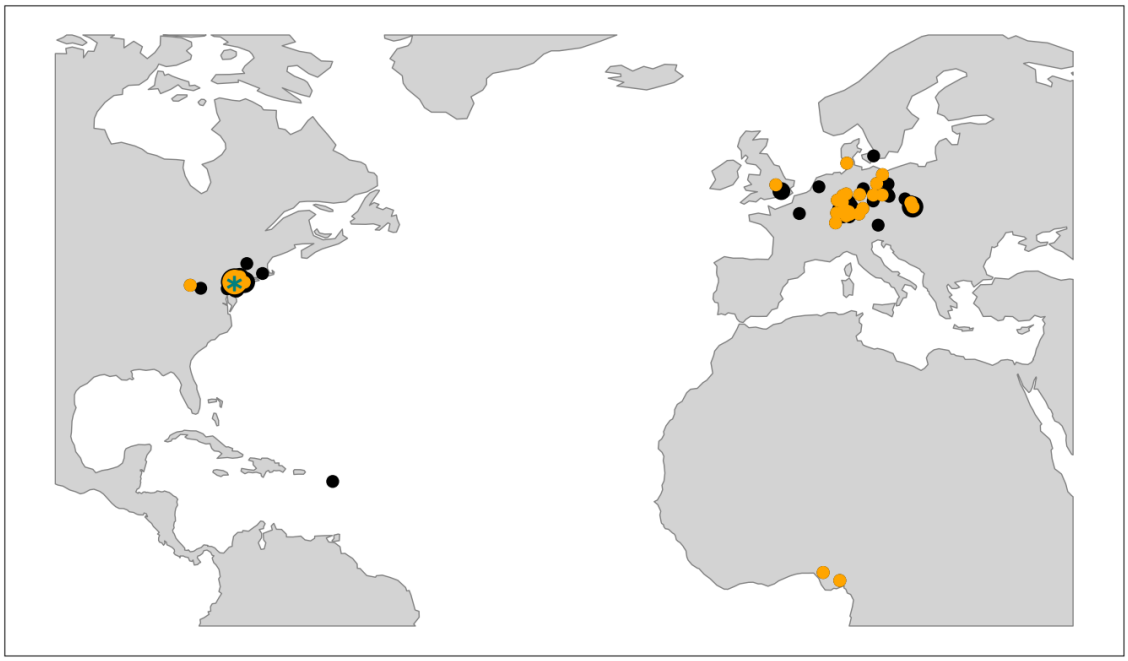

Figure 5.1: Birth places of Moravians with memoirs considered here (all 89 in black, the 36 ones selected for manual annotations in orange, the dot size reflects the amount of memoirs; the teal asterisk marks the place where the texts were archived)

There are two main reasons why not all areas of birth are represented in the subsample: First of all, the process still included a randomization step, and, secondly, there is the second balancing effort regarding the gender distribution. The latter prevents the subsample to be annotated from overrepresenting girls and women as a result of the sampling precedence of the original digitization projects (see Section 4.1.3). The combination of these subsampling targets resulted in the selection of 18 memoirs per gender or 36 in total (see Table 5.1).

Dataset	**Biographies**		
	Female	**Male**	**Total**
BDHP	6	11	17
bethlehemer-schwestern	35	–	35
moravian-mens-memoirs	–	5	5
moravian-lives_de	13	19	32
sample	54	35	89
subsample	18	18	36

Table 5.1: Gender distribution in Moravian dataset subsamples

Based on the representation of these diverse sociolinguistic backgrounds, the subsample can be used on its own to conduct fine-grained analyses. This is even more true if enriched with more texts in the future. The debiasing attempts also increase the adequacy of the subsample for the evaluation of different automatic annotation approaches with regard to this specific domain and time.

Sentence tokenization

As a first preprocessing step for future annotation tasks, all 36 memoirs in the subsample were split into sentences with the Python module *stanza* (Qi et al., 2020). The sentence tokenization failed in a few cases: One issue involved incomplete phrases that ended with ordinal numbers because the original sentence contained a date. These cases were automatically identified with regular expressions and merged with the subsequent instance. In the case of Example (5.1), stanza

split the sentence after the abbreviation of a month name (= *Febr.*) but the resulting two sentences were automatically merged afterwards.

(5.1) *Er ward geboren den 24ten Febr.*
1725 zu Zauchtenthal in Mähren.
('He was born on the 24th of Febr.
1725 in Zauchtenthal in Moravia.')

Another problem can be attributed to punctuation which is not always consistent with contemporary German rules as expectable from 18th century sources which were written before the emergence of such rules. More specifically, sentences in a linguistic sense were found to not only be separated by dots, question and exclamation marks but also by comma, semicolon, or a simple blank in these sources. The latter three sometimes led to rather long sentences. These were manually split as shown in the following Example (5.2) which was a single sentence in the stanza output but later split into two sentences:

(5.2) *Der Heiland aber war so treu und lies ihr keine Ruhe, bis Sie wieder auf ihr Herz kam, und über alles Sünderin wurde,*
Sie bereuete auch hernach die Zeit oftt mit vielen Thränen, und bat es dem Heiland ab.
('The Savior, however, was so faithful and gave her no peace until she found her heart again and became a sinner above all else,
She often regretted thereafter the time with many tears and begged the Savior for it.')

The reason for focusing manual preprocessing efforts on instances like the two previous examples was that both underlying sentence tokenization issues can be identified rather straightforwardly and are also problematic for the actual annotation tasks at hand. More specifically, the falsely split second "sentence" in Example (5.1) lacks the verb which is highly relevant for the life event described. Similarly, a falsely merged sentence, such as Example (5.2), is composed of two or more individual sentences in a linguistic sense which deal with different life events and these events may potentially be portrayed with different polarities and occur during different life stages. This is why an unsplit "sentence" of that kind is more likely to require a later disambiguation step than a shorter sentence. All these preprocessing steps led to a final corpus comprising 2210 instances (see Table 5.2).

Dataset	**Memoirs**	**Sentences**	
		stanza	**cleaned**
BDHP	17	1893	–
bethlehemer-schwestern	35	2301	–
moravian-mens-memoirs	5	773	–
moravian-lives_de	32	1506	–
sample	89	6473	–
subsample	36	2361	2210

Table 5.2: Number of sentences in Moravian dataset subsamples

Anonymization

As a final preprocessing step, each name of a person was manually masked with the abstract token `{NAME}` in each one of the 2210 sentences (Brookshire and Reiter, 2024, 92). This was deemed necessary due to the potential ethical risk of training machine/deep learning models to base their algorithmic decisions on personal names. This kind of debiasing was proposed by Hube et al. (2020) because such a model could be applied to other datasets in the future where such a bias poses an even more severe ethical risk (see Section 3.1.4).

Chapter 6

Sentiment classification

The subsample of the German Moravian memoirs outlined in the previous Chapter 5 was manually annotated with sentiment labels as described in Section 6.1. In addition, various automatic annotation approaches mentioned in related work were selected based on criteria which are discussed in Section 6.2. The accordance between the manual annotations on the one hand and automatic ones on the other are presented in Section 6.3.

6.1 Manual annotation

Section 3.1.2 hinted at the plethora of resources and tools for sentiment classifications which are already available off-the-shelf. However, as outlined in Section 3.1.4, sentiment annotations tend to have a rather high sensitivity towards the target domain and language variety at hand. This is why a novel manually annotated dataset was introduced by Brookshire and Reiter (2024, 92–93) in order to conduct sentiment analyses within this thesis. It is, to the best of my knowledge, the first one explicitly targeted at German autobiographical sources from the 18th and 19th century (see Section 5). It is annotated on the sentence level based on the principles of a ternary sentiment analysis defined in Section 3.1.1. In practice, this means that the label `negative` is used whenever the main event described in a given instance is portrayed (rather) negatively and the label `positive` in the opposite case. The `neutral` class is only attributed to *truly neutral* sentences – i.e. ones which lack any kind of sentiment-related information. The exact label definitions are provided in Table A.1 in the Appendix. Table 6.1 lists one example per class with translations approximating peculiarities in the original casing and punctuation.

Label	Text	Translation
`negative`	Darüber wurde sie böse und ging davon.	She got angry about this and left.
`neutral`	1760. zog sie nach Naz.	In 1760. she moved to Naz.
`positive`	er war ein guter Mensch,	he was a good person,

Table 6.1: Examples from the manually annotated Moravian memoirs per sentiment class

Whenever a sentence contained words with opposing polarities, the label attribution was resolved by only considering the "final state or result of the action described" in the given instance (Brookshire and Reiter, 2024, 93). As discussed in the paper, Example (6.1) was attributed the label `negative` as it ends on a darker note even though the first part of the phrase involves a certain degree of positivity.

(6.1) *Ich versuchte oft und viel mir selbst aus diesem Zustand zu helfen, aber vergebens,*
('I tried often and hard to help myself out of this state, but in vain,')

This strict class definition was employed to increase the intersubjectivity of the annotations despite the fact that only one single (expert) annotator was feasible in the given project setting.

Nevertheless, especially the strictness with regard to neutral instances may be part of the reason why most instances were considered non-neutral with a slight bias towards the label `positive` (see Table 6.2). These findings are, however, also relatable to the expected high degree of emotionality present in Moravian memoirs as a textual genre (Van Gent, 2017; Faull and McGuire, 2022).

Label	**Dataset**		
	Train	**Test**	**Total**
`negative`	485	115	600
`neutral`	523	150	673
`positive`	**760**	**177**	**937**
Total	1768	442	2210

Table 6.2: Statistics of Moravian sentiment datasets (highest support in bold; based on Brookshire and Reiter 2024, 93)

Table 6.2 shows that the distribution between classes (i.e. 40–43 % are `positive`, 30–34 % `neutral` and 26–27 % `negative`) stayed quite consistent after randomly assigning sentences to the train and test datasets. The split was implemented with the Python module *datasets* (Lhoest et al., 2021) as is common practice. It resulted in 1768 training instances (= 80 %) and 442 (= 20 %) hold back for testing. While neither a dataset split nor a sentence-wise randomization are necessary for analyzing the dataset on its own, they are deemed important to (i) increase the generalizability of systems trained on this data and (ii) to prevent information leakage.

6.2 Automatic annotation

Considering the amount of digitally available Moravian memoirs, it is feasible to annotate all of them manually. However, since the amount is expected to increase (see Section 4.1.3), it is also worth assessing automatic annotation approaches in order to facilitate an upscaling as soon as more texts become digitized. The different methods evaluated within this thesis range from sentiment dictionary lookups (Section 6.2.1) over "traditional" machine learning approaches represented by SVMs trained on the manual annotations (Section 6.2.2) to transformer models (Section 6.2.3). The latter include off-the-shelf models as well as ones fine-tuned on the manual annotations. Finally, a few zero-shot models (Section 6.2.4) were selected which do not require any kind of task-specific training data.[20]

6.2.1 Dictionary-based approaches

Dictionary look-ups are not only rather straightforward to implement but also quite common in digital humanities projects in general and in ones dedicated to biographical sources in particular. This is why they are the first automatic annotation approaches to be presented here. There are many different sentiment dictionaries in use (see Section 3.1.2). However, some of them actually assign emotion instead of sentiment polarity labels and are hence not considered here. One such example is SentiArt which had the highest coverage with regard to a dataset from a similar time frame (see Table 3.1). Yet, other dictionaries listed in the respective study (Herrmann and Grisot, 2022, 168) are able to assign sentiment labels. They were hence used as a starting point for the selection of this work. Further ones were added from the surveys of Fehle et al. (2021, 89) and Kern et al. (2021, 37:4–5). In order to also account for different compilation strategies, sizes (i.e number of entries), and intended target domain, the following eight German sentiment lexicons were selected:

AffectiveNorms (Köper and Schulte im Walde, 2016). With ratings for more than 350,000 word forms in four affective dimensions (only `valence` is considered here), it is currently the largest German sentiment resource available. The ratings were inferred with a supervised learning algorithm trained on three other sentiment lexicons (with BAWL-R being one of them).

[20] Section 2.3.1 provides more general information about the different systems.

BAWL-R (Võ et al., 2009). The *Berlin Affective Word List Reloaded* is one of the earliest German sentiment resources. It lists values for almost 3000 word forms in three affective dimensions and ten psycholinguistic indexes (only `emo[tional valence]` is considered here). The values are averaged from manually assigned scale ratings by 200 different psychology students in total.

Germanlex (Clematide and Klenner, 2010). This semi-automatically compiled sentiment lexicon lists polarity ratings for almost 9000 word forms. It is derived from a list of adjectives automatically extracted from newspapers which are likely to bear sentiment information. It was compiled with the target domain of German novels in mind.

GermanPolarityClues (Waltinger, 2010b). This dictionary used to list ternary sentiment labels for 10,000 German lemma forms but was later extended to account for almost 38,000 (inflected) word forms in total. It was compiled by automatically translating English dictionaries enriched by synonyms and negated phrases. Some of the annotations were manually assessed and corrected.

GerVADER (Tymann et al., 2019). This German adaptation of the *Valence Aware Dictionary for sEntiment Reasoning (VADER)* (Hutto and Gilbert, 2014) uses contextual rules in order to factor in degree modifiers and sentiment shifters. As both language versions were developed to analyze social media data, they consider features like punctuation and capitalization which may be less predictive here. This is even more true for the underlying lexicon which differs from SentiWS mainly only due to the inclusion of a few slang words.

MultilingualSentiment_de (Chen and Skiena, 2014). This multilingual dictionary covers 136 languages in total and lists binary sentiment labels for almost 4000 German word forms. The labels were propagated from existing English sentiment lexicons with a multilingual knowledge graph based on semantic links between individual word forms. The link types include translation equivalents, equal spelling in closely related languages, and synonyms.

SentiMerge (Emerson and Declerck, 2014). With almost 98,000 word forms, this polarity dictionary is the second largest in the selection. It was derived from four existing German sentiment lexicons (including smaller versions of Germanlex, GermanPolarityClues, and SentiWS) using a Bayesian probabilistic model.

SentiWS (Remus et al., 2010). This is of the most widely used German dictionaries with polarity values. It includes more than 34,000 word forms and focuses on the financial domain. The initial seed list of lemma forms were automatically translated from an English dictionary and subsequently enriched by co-occurrence analyses of rated product reviews as well as a collocation dictionary. The list of lemma forms was automatically extended to inflected ones.

Coverage when applied to Moravian memoirs

The intended target domain is not always as clearly described in the introductory paper of all dictionaries. Nevertheless, most target domains can be expected to be quite distinct from 18th century biographical sources. In order to verify this, a coverage analysis was conducted with the results being summarized in Table 6.3. Unlike the aforementioned coverage analysis (see Table 3.1), this one also assesses the potential impact of further preprocessing steps, such as casing or lemmatization. The latter was implemented with stanza. The coverages with regard to Moravian memoirs are in a rather low range similar to the one observed by Herrmann and Grisot (2022, 168). In fact, the mean cased token coverage of 15 % is even lower than theirs (= 21 %). This is partly a result of not considering the emotion dictionary SentiArt with its 82 % coverage on their data. Interestingly, the distributions seem to hint at a somewhat expectable correlation between dictionary size and coverage. This was not the case in the analysis of Herrmann and Grisot (2022) where the rather small BAWL-R had one of the highest coverages. However, it is still more of a slight tendency as Germanlex yields slightly lower coverages than BAWL-R here despite listing allmost three times as many word forms.

It should be highlighted that longer dictionaries like AffectiveNorms and SentiMerge appear to yield more net gains when applied to uncased and/or lemmatized texts. This is somewhat counterintuitive as both preprocessing steps reduce the number of distinct word forms which should in turn increase the likelihood that a given word form is found in a given dictionary. This finding might relate to the domain issue albeit that novels (e.g. Germanlex) may be expected to share

more similarities with autobiographical sources than financial data (e.g. SentiWS). However, the smaller dictionaries profit only marginally from the two preprocessing steps with lemmatization even reducing the coverage of GerVADER and SentiWS. In summary, these mixed results do not really support the inclusion of such preprocessing steps.

Lexicon	Size		Token coverage		Lemma coverage	
	cased	uncased	cased	uncased	cased	uncased
AffectiveNorms	**351,616**	**345,416**	**.32**	**.39**	**.52**	**.55**
BAWL-R	2,902	2,902	.08	.09	.14	.15
Germanlex	8,691	8,644	.07	.08	.12	.13
GermanPolarityClues	37,589	37,367	.16	.18	.17	.18
GerVADER	34,585	34,380	.13	.14	.10	.11
MultilingualSentiment_de	3,974	3,891	.08	.09	.12	.12
SentiMerge	97,828	96,988	.27	.31	.48	.50
SentiWS	34,318	34,125	.13	.14	.10	.11

Table 6.3: Sizes of common German sentiment dictionaries and their coverage when applied to the 36 manually annotated Moravian memoirs (highest values in bold)

Normalization efforts

Since all selected dictionaries but GerVADER operate on the token instead of the sentence level, an aggregation strategy was necessary. This was implemented by averaging the dictionary values of all tokens in a given sentence without further rule-based weightings like in the case of (Ger)VADER. Hence, it is possible that a different strategy, such as one which ignores word forms not found in a given dictionary as proposed by Chen and Skiena (2014, 386–387), yields better results.

The comparability between the selected dictionaries was further increased by rescaling the values to the common interval `[-1,1]` as proposed by Kern et al. (2021, 37:3). As some lexicons are imbalanced between the two polar classes, each value was divided by the maximum absolute one in order to maintain the imbalance. This affected the values of BAWL-R with a default range of `[-3,2.9]` mapped to `[-1,.97]`, GermanLex with `[-7,5]` mapped to `[-1,.71]`, and SentiMerge with `[-1.63,1.52]` mapped to `[-1,.93]`.

6.2.2 SVMs

SVMs belong to the earliest systems used for sentiment classification in the early 2000s (Pang et al., 2002) and have ranked amongst the most performant ones in many studies in the two decades thereafter (Waltinger, 2010a; Sidarenka, 2017; Hankar et al., 2025). One of the first SVM-based sentiment classifiers was proposed by Pang et al. (2002) and showed that feature vectors solely based on binary token presence yield the best result. Davoodi and Mezei (2022) recently observed that state-of-the art performances competitive to the ones of transformer models are reachable by leveraging TF-IDF weightings. Another promising alternation seems to be the use of character level n-grams instead of tokens (Mohammad et al., 2013). Based on these recommendations, the following three parameter settings of a linear SVM fitted on distributional features of the Moravian train dataset will be compared here: (i) **different vectorizers** (= binary presence, TF, and TF-IDF), (ii) **different n-gram ranges** from uni- to trigrams and (iii) **different analyzer levels** (= whole words and individual characters). The latter two parameters determine the dimensionality of the vectorization the SVMs are based on (see Table 6.4).

The dimensionalities imply that the Moravian dataset is less diverse than the one of Pang et al. (2002, 83) as he proposed using 16,000 features (= the number of words occurring at least 4 times in their corpus) while the total number of different word unigrams is only 6000. Considering their second best setting (= a combination of unigrams and bigrams), the number of features are quite similar when again relying on the full amount here. Hence, not using a cut-off seems to be feasible

in this lower resource setting. In addition, it has the advantage of reducing out-of-vocabulary issues during inference.

N-gram range	Analyzer	
	word	char
1	6,168	75
1–2	33,675	1,252
1–3	71,270	7,932
2	27,507	1,177
2–3	65,102	7,857
3	37,595	6,680

Table 6.4: Number of SVM feature dimensions by n-gram range and analyzer when vectorizing the Moravian train dataset

Finally, it should be noted that the actual vectorization and fitting of SVMs with a linear kernel was implemented with the widely adopted Python module *sklearn* (Pedregosa et al., 2011). This took 2.5 minutes on a regular CPU in total for all 36 resulting settings.

6.2.3 Transformer models

Transformer models have been shown in various surveys as well as individual studies to currently be the most performant sentiment classification systems – even more so when fine-tuned to a specific text domain or language variety (Guhr et al., 2020; Alaparthi and Mishra, 2021; Corvonato, 2021). Since sentiment analysis is one of the most widespread text classification tasks, it is not so surprising that there are quite a lot of models readily available on the Hugging Face Model Hub. A simple search for *sentiment* yielded almost 12,000 results at the End of January 2026 and one restricted to the German language still 89. Obviously, it is not feasible to evaluate all of them which is why the number of downloads was chosen as a proxy for widespread usage. In order to focus on fine-tuning, models distilled from a zero-shot classification pipeline were not considered. This preselection resulted in three models with considerably more downloads (= 664,000, 260,000, and 96,800) than the fourth one (= 16,600). As the first two were both multilingual ones but also showed a notable download difference, only the first and third one were chosen:

german_sentiment_bert (Guhr et al., 2020). This model was fine-tuned on 1.8 million instances of social media and customer review data and assigns ternary sentiment labels.[21]

multilingual_sentiment_bert (NLP Town, 2023). This model was fine-tuned on 629,000 product reviews in six languages (the German part comprising 137,000). It predicts star ratings between 1 and 5. In order to increase the comparability, these ratings were mapped linearly to polarity values in the interval `[-1,1]` (see Section 6.2.1).

Transformer fine-tuning

Laurer et al. (2024, 95) suggested fine-tuning a BERT model whenever 1000 or better 2000 training samples are available for a task at hand. Hence, both aforementioned BERT models were fine-tuned for a second time with the 1768 Moravian sentences manually annotated with sentiment labels (see Section 6.1). In addition, the following three base models, which were not previously trained to conduct a sentiment classification, were also fine-tuned with the same sentences:

bert (Devlin et al., 2019). The cased version of the original English BERT model from Google pre-trained on approximately 12 GB of books and the English Wikipedia. It is primarily included as a kind of lower baseline but also to test its cross-lingual transfer learning ability. It may benefit from the place of origin of the training data (see Figure 5.1) which may have led to English loans.

bert_multilingual (Devlin et al., 2019). The cased version of this multilingual equivalent was pre-trained on texts in the 104 languages with the largest Wikipedia. It may also benefit to some

[21] There is the Python module *germansentiment* but *transformers* was used as in the case of the other BERT models.

degree from the fact that not all Moravians the fine-tuning (and inference) data is about were native speakers of German.

gbert (Chan et al., 2020). The cased version of a German equivalent from deepset pre-trained on approximately 163 GB of data crawled from German websites. Yet, it is quite unlikely that it has seen Moravian sources during its original sentiment fine-tuning step.

All five BERT models were fine-tuned over three epochs using default parameters as suggested by Devlin et al. (2019). The fine-tuning was implemented with the Python module *transformers* (version 4.28.0) on a T4 cloud GPU (Brookshire and Reiter, 2024) as is common practice. It took approximately nine minutes per model supporting the observation of Laurer et al. (2024, 69) that such a transfer learning setting is more resource-intensive than an SVM fitting.

6.2.4 LLMs and other zero-shot approaches

The last group of automatic annotation approaches evaluated here are zero-shot approaches. They have been shown to outperform dictionary-based ones but underperform compared to SVMs or transformer models trained/fine-tuned on the task and/or domain at hand (see Section 3.1.2).

LLM selection

In the case of sentiment analysis, LLMs have been shown to yield promising results in a zero-shot setting although their "true capacity [...] remains unclear" according to Zhang et al. (2024, 3883). Part of the reason for this slight contradiction is the problem that the capacity or performance is often evaluated on datasets which are publicly available and therefore potentially already part of the data an LLM was trained on. Such *information leakage* effects apply to sentiment analysis in particular because it is arguably the most common text classification task. It is hence not surprising that only 8 (= 30.8 %) out of 26 sentiment datasets are attributed a low leak severity in a recent survey (Balloccu et al., 2024, 72). However, information leakage is an even more challenging problem when presupposing a zero-shot setting. More specifically, LLMs may adapt their predictions to prior user interactions which obfuscates the actual number of shots involved. For the sake of reproducibility and to at least avoid future information leakage, only LLMs were considered that (i) run locally on institutional servers and (ii) do not store or cache data for future model improvements. The following two models were selected:

Gemma3_27B (Gemma Team, 2025). This multimodal model with multilingual capabilities developed by Google DeepMind is based on a decoder-only transformer model. It has a context size of 128,000 tokens and 27 billion parameters. It is instruction-tuned with a post-training focused on "improving mathematics, reasoning, and chat abilities".

Nemotron_Ultra_253B (Bercovich et al., 2025). This reasoning model developed by NVIDIA is a distilled derivative of a Llama-3 model with 253 billion parameters. It has a context size of 128,000 tokens. Its post-training included supervised fine-tuning as well as reinforcement learning steps all targeted at improving its reasoning capabilities in domains like Math, Coding, and Science.

Prompt templating

The most important features of the prompt design were two-fold: First of all, the prompt template needs to be generalizable to all classification tasks conducted within this thesis in order to also enable a cross-task comparison. Secondly, the prompt should lead the LLMs to only elicit a single class label since they are applied to text *classification* tasks instead of their more prototypical text *generation* or reasoning abilities. With regard to the taxonomy suggested by Liu et al. (2023, 195:10), this means that the desired **shape** of an answer `[Z]` is the token granularity and the desired **answer space** a constrained one. This is reflected in the basic prompt template used for all three classification tasks of this thesis:

(6.2) *Please perform a [C] classification task. Given the sentence, assign a [C] label from [Y]. Return label only without any other text. Sentence: [X] ... Label: [Z]*

The prompt template shown in (6.2) is based on the one Zhang et al. (2024, 3884) used for sentiment analysis. It is, however, more general in the sense that *sentiment* was replaced with the generic task name `[C]` and the answer space constraining list of class labels `[Y]` was adapted accordingly. It should be noted in this context that the latter is also true for the original study as Zhang et al. (2024, 3898–3900) analyzed not only four binary sentiment datasets but also one ternary and two with a 5 item scale. The final sentiment prompt template is:

(6.3) *Please perform a sentiment classification task. Given the sentence, assign a sentiment label from ["negative", "neutral", "positive"]. Return label only without any other text. Sentence: [X] ... Label:*

The input sentence `[X]` was added during inference for which the API of the locally hosted models and not a web interface was used. It should also be mentioned that a mapping to the final output is necessary whenever the answer `[Z]` of the model is not part of the list of labels `[Y]` (Liu et al., 2023, 195:10). This mapping encompassed only the removal of line breaks sometimes found in the answers of Gemma3_27B but no further postprocessing.

Selection of further zero-shot models

Besides the two LLMs, the three base transformer models bert, bert_multilingual, and gbert were also evaluated with regard to their zero-shot capabilities. In addition, the following two NLI models were considered since Zhang et al. (2025) stated that this approach is currently the second most common one in the context of zero-shot text classification:

bart-nli (Lewis et al., 2020). The *Bidirectional and Auto-Regressive Transformer (BART)* is a denoising autoencoder with 406 million parameters. It was trained on a crowd-sourced collection of 433,000 English NLI sentence pairs.

mDeBERTa-nli (Laurer et al., 2024). The *multilingual decoding-enhanced BERT with Disentangled Attention (mDeBERTa)* has 279 million parameters and was fine-tuned on almost 3.3 million hypothesis-premise pairs in 27 languages (including German). It yielded micro F1 scores between .46 and .76 when applied to humanities datasets (Klähn et al., 2025, 8).

Since both models are based on the transformer architecture, the automatic annotation was implemented with the *zero-shot classification pipeline* of the Python module *transformers* (version 4.28.0) as is common practice.

6.3 Evaluation

All systems mentioned in the previous section are compared to each other with regard to their accordance to the manual annotations in Section 6.3.1. Section 6.3.2 then discusses notable classification errors of the best performing systems before the respective classification behavior is analyzed even further with explainability approaches in Section 6.3.3. All these different evaluations aim at an assessment of whether the selected automatic annotation approaches allow for an upscaling or if they fail to grasp domain-specific peculiarities.

6.3.1 Performance comparisons

In addition to the normalization steps described in Section 6.2.1, the polarity values gained from the dictionary-based approaches as well as multilingual_sentiment_bert needed to be discretized to the three sentiment labels. This was achieved by a mapping based on the threshold of `±.05` as proposed by Hutto and Gilbert (2014). More specifically, the interval `[-1,-.05]` was mapped to the label `negative`, `[-.05,.05]` to `neutral`, and `[.05,1]` to `positive`.

The actual evaluation was based on the three common scores **precision**, **recall** and **F1 score** as calculated with the Python module *sklearn*. A weighted averaging was used and the results are summarized in Table 6.5. Besides the different systems, the table includes a random and majority baseline for reference. The latter always assigned the label `positive`. In the case of SVMs, only

the best result per analyzer is reported while the results of all 36 evaluated parameter settings are summarized in Figure 6.1. LLMs were evaluated in five runs on five different days in order to account for non-determinism. The table lists the mean and standard deviation of these runs.

System	Type	Precision	Recall	F1 score
Random	baseline	.34	.34	.34
Majority	baseline	.16	.40	.23
AffectiveNorms	dictionary	.32	.42	.36
BAWL-R	dictionary	.53	.37	.28
Germanlex	dictionary	.12	.34	.17
GermanPolarityClues	dictionary	.50	.47	.47
MultilingualSentiment_de	dictionary	.43	.40	.38
SentiMerge	dictionary	.12	.34	.18
SentiWS	dictionary	.29	.34	.18
GerVADER	dictionary + rules	.59	.59	.58
SVM_char2-3_TF-IDF*	SVM*	.70	.69	.69
SVM_word1-2_TF-IDF*	SVM*	.72	.71	.70
bert*	BERT*	.65	.63	.63
bert_multilingual*	BERT*	.77	.76	.76
gbert*	BERT*	.81	**.81**	**.81**
german_sentiment_bert	BERT	.56	.37	.25
german_sentiment_bert*	BERT*	.74	.74	.74
multilingual_sentiment_bert	BERT	.41	.47	.42
multilingual_sentiment_bert*	BERT*	.78	.77	.77
bert_110M	zero-shot (BERT)	.34	.34	.34
bert_multilingual_110M	zero-shot (BERT)	.40	.39	.38
gbert_110M	zero-shot (BERT)	.37	.36	.35
bart-nli_406M	zero-shot (NLI)	.53	.38	.33
mDeBERTa-nli_279M	zero-shot (NLI)	.48	.53	.47
Gemma3_27B	zero-shot (LLM)	$.76_{\pm.01}$	$.75_{\pm.01}$	$.75_{\pm.01}$
Nemotron_Ultra_253B	zero-shot (LLM)	$\mathbf{.82}_{\pm.00}$	$.77_{\pm.00}$	$.77_{\pm.00}$

Table 6.5: Sentiment classification results (best value per score in bold, systems fine-tuned/trained on Moravian data marked with *, only the best results per analyzer are listed in the case of SVMs)

In general, the scores lead to a ranking between the different system types which is in line with the ones reported in related work: BERT models performed best with scores ranging between .25 and .81 (mean .65), and were followed by SVMs fitted on distributional features with scores in the range .44–.72 (mean .61). Zero-shot models ranked third (.33–82, .51), and dictionary-based systems last (.12–.59, .37). Furthermore, systems adapted to the subject matter by being trained on the manual annotations outperformed all off-the-shelf systems but the two evaluated LLMs by a margin. Hence, most off-the-shelf systems seemingly struggled with the domain transfer from mostly modern-day German (as found in social media data and customer reviews) to the Moravian sources under investigation. More detailed observations worth mentioning are inferable from the scores alone when zooming in on individual systems.

Performance of dictionary-based approaches

Whereas all three performance scores are quite similar for most systems, dictionary-based ones had more issues with regard to their *precision* (which also applies to the majority baseline). Germanlex and SentiMerge, for instance, yielded only a precision of .12 which is not only 23 % points less than their recall but also the lowest score in the whole system comparison. Still, it is a mere tendency

as BAWL-R, by contrast, had the second highest precision of all evaluated dictionaries (.53) but a poorer recall (.37). GermanPolarityClues as well MultilingualSentiment_de yielded a slightly higher precision than recall, too. The high precision of BAWL-R is most likely related to the fact that its values are fully based on manual ratings and thus the highest quality of all considered dictionaries. The other two dictionaries with a higher precision have in common that they assign sentiment labels instead of finer-grained polarity values. This seems to work better in the simple averaging approach followed here as they outperform dictionaries of similar sizes by noteworthy F1 score margins. For instance, MultilingualSentiment_de beat BAWL-R by 10 % points and GermanPolarityClues SentiWS even by 30 % points. One possible reason for this difference may be that polarity values weighted between 0 and 1 are too low to grasp a more implicit sentence-wise orientation even with the rather low threshold used.

In fact, all scores are pretty low and neither obviously correlated to the respective dictionary size nor their coverage (see Table 6.3). The most notable "misfit" was SentiMerge with its second highest size and token coverage but second worst performance. The only promising alteration appears to be the inclusion of contextual rules which factor in intensifiers and negations as in the case of GerVADER which not only beat the F1 score of the second best dictionary (= GermanPolarityClues) by an 11 % point margin but also SentiWS by 40 %. Since contextual rules are the main difference between GerVADER and SentiWS, this finding is in line with the claim of Kim and Klinger (2021) that a lack of context-awareness is a major flaw of dictionary-based sentiment analysis but can be circumvented to some extent when including such rules. Hence, it cannot be ruled out that the best performing dictionary may also yield better scores with such rules although it is unlikely that it overcomes the 19–23 % point F1 score difference towards the three best performing systems.

Performance of SVMs

Since an SVM fitting requires rather few computational resources when using approximately 2000 samples for training, different vectorization approaches were evaluated as well (see Section 6.2.2). While only the most performant setting per n-gram analyzer is listed in Table 6.5, the effect of different n-gram ranges and vectorizers are displayed in Figure 6.1.

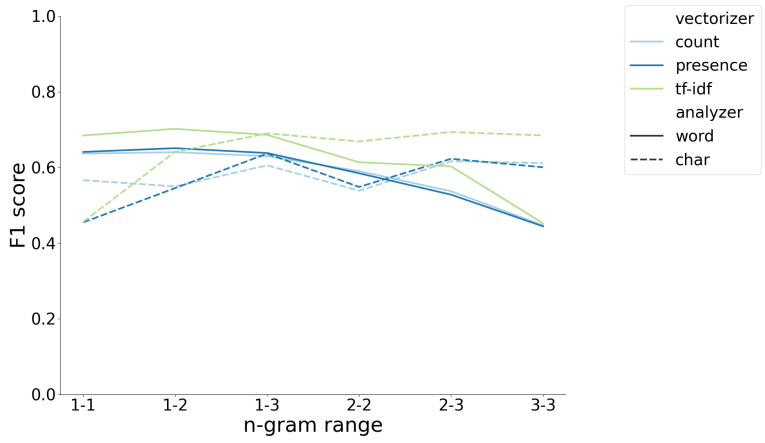

Figure 6.1: Influence of different parameter settings on SVM perfomances

The figure shows that TF-IDF weightings outperformed the two other simpler ones no matter the n-gram range. However, unlike in the study of Pang et al. (2002), a binary n-gram presence vectorization yielded similar results to raw frequency counts. A combination of character bi- and trigrams performed almost as well as their proposed combination of word uni-bigrams despite the by far lower dimensionality (see Table 6.4). Yet, both granularities behaved different when altering the n-gram range no matter the vectorization type. More specifically, the performance of word unigrams was only slightly improved by including bigrams but worsened after the addition of trigrams or removal of unigrams. Character n-grams, by contrast, definitely profited from the additional "context" of bi- and trigrams.

Performance of fine-tuned transformer models

All selected off-the-self transformer models yielded rather poor performances which are comparable to the four most performant dictionary lookups. This fits the observation of Klähn et al. (2025, 8) that both system types tend to have out-of-domain issues. The fine-tuned versions, by contrast, beat the most performant SVM settings by 4–12 % points and gbert the evaluated LLMs by 4 % points. The weighted F1 score of .81 of the fine-tuned gbert can be seen as "state-of-the-art" as Schmidt et al. (2022), for instance, reached similar values with German BERT models applied to German drama texts from a similar time frame.

The net gains of fine-tuning an off-the-shelf sentiment model were 49 % points in the case of german_sentiment_bert and 35 % points for the multilingual model. As a matter of fact, these gains surpass the effect of adding contextual rules to SentiWS. Therefore, one could come to the conclusion that transformer models seemingly keep their ability to adapt to small datasets even when they have already been fine-tuned on a similar task with more data before.

Interestingly, the F1 score of the fine-tuned multilingual_sentiment_bert (= .77) is only slightly better than the one of the basic bert_multilingual (= .76) which lacks a previous fine-tuning step. Regarding this fine-tuning setting, it is somewhat expectable that the German base model performed best whereas its multilingual counterpart performed a little worse and the English one worst. Still, it should be highlighted that even the fine-tuned English base BERT model outperformed not only all dictionary-based approaches but also transformer models fine-tuned for contemporary German sentiment by a margin. In summary, this implies that, in a low-resource setting, it may be better to fine-tune a base model directly than first training a task-specific off-the-shelf model for a different domain and later adapting it to the target domain. This is even more true given that a single fine-tuning step obviously requires less computational resources.[22]

Performance of zero-shot approaches

In the case of zero-shot approaches, a parameter count effect is observable since the model with the highest amount of parameters (= Nemotron_Ultra_253B) performed best and Gemma3_27B second best. However, the net gain of adding almost ten times as many parameters, which is one difference between the two selected LLMs, was actually minor. The same is true for the parameter size differences between the smaller transformer models with less than half a billion parameters. As the main performance difference of a 28–44 % F1 score margin was between these two groups, this can be seen as a potential emergent feature of LLMs.

Considering the smaller models, mDeBERTa-nli underperformed compared to transformer models fine-tuned on data from the target domain but still outperformed base BERT models in a zero-shot setting as predicted by the developers (Laurer et al., 2024, 95). The basic English BART model fine-tuned for NLI, by contrast, was only able to surpass base BERT models in the case of precision despite having significantly more parameters and requiring more resources.

In conclusion, only the two LLMs yielded performance scores similar to the ones of fine-tuned transformers. This implies that they may be useful in the case of a ternary sentiment analysis despite their high demand of resources at inference time. They also seem to assign labels in more stable manner than expectable as underlined by the rather low standard deviations.

6.3.2 Error analyses

A first step towards a deeper understanding of the behavioral differences between the selected automatic sentiment annotation systems was achieved by looking at **confusion matrices**. These are depicted for the best performing dictionary (= GermanPolarityClues), dictionary with rules (= GerVADER), SVM setting (= TF-IDF weighted word uni- and bigrams), off-the-shelf and fine-tuned transformer models (= multilingual_sentiment_bert and gbert), and best LLM run (= the fifth of Nemotron_Ultra) in Figure 6.2.

[22] A similar consideration was the reason why the reverse transfer learning strategy of first exposing a base BERT model with unlabeled Moravian data before fine-tuning it with sentiment labels was not evaluated.

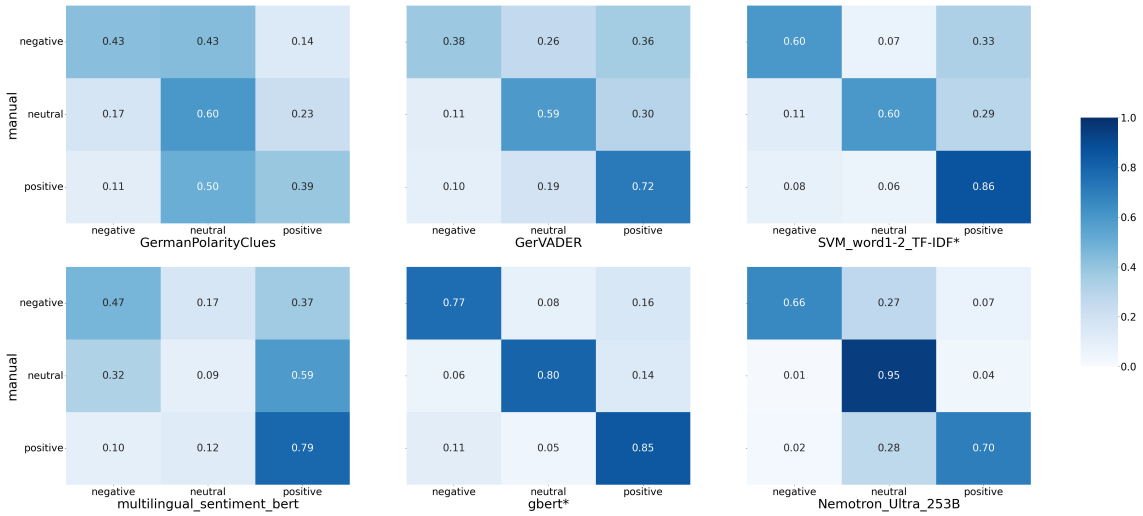

Figure 6.2: Confusion matrices of best performing sentiment systems per type (systems fine-tuned/trained on Moravian data marked with *)

The confusion matrices illustrate that neither one of the six systems assigned sentiment labels in a random manner. Instead, GermanPolarityClues and Nemotron had a bias towards the label `neutral` which is more pronounced in the case of the dictionary than the LLM. In the former case, this supports the hypothesis from the previous Section 6.3 that either the polarity weightings of some of the dictionaries may be too low or the threshold too high to be more in line with the manual annotations. In the case of Nemotron_Ultra, it is worth mentioning that no other system predicted more neutral sentences correctly. In addition, it confused less negative samples with positive ones than gbert despite the 4 % point F1 score difference. Hence, the main difference between both systems is that said LLM had issues distinguishing polar sentences from neutral ones while gbert had more issues in correctly predicting the actual polarity.

Being dictionary-based, it is not so surprising that GerVADER also shows a slight tendency towards the `neutral` class because coverage issues may arise. However, it also appears to favor the label `positive` which is in line with the respective remark in its introductionary paper (Tymann et al., 2019, 87–88). Furthermore, GerVADER performed better than its source dictionary SentiWS which basically classified almost all sentences as `neutral` (see Figure 6.3).

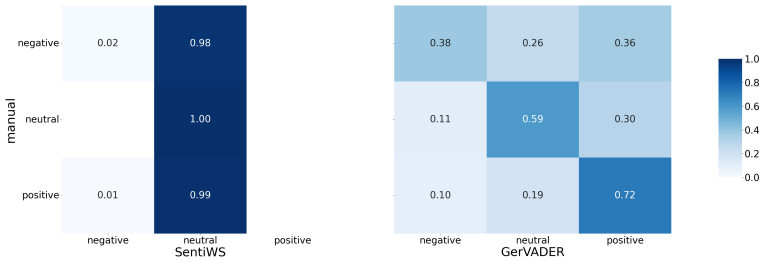

Figure 6.3: Influence of adding rules based on the confusion matrices of SentiWS and GerVADER

The confusion matrix of the off-the-shelf transformer model multilingual_sentiment_bert stands out because it implies that this system struggled most with the identification of truly neutral sentences. The same peculiarity was only also observed in the case of the slightly better performing zero-shot model mDeBERTa-nli. Hence, it may hint at a general bias of such models towards one of the two polar classes with a slight preference for a `positive` output which is in line with findings from related work (see Section 3.1.1).

In the case of the multilingual sentiment model, the low performance with regard to neutral instances may partly be a side effect of the necessary preprocessing step. More specifically, the polar classes account for two of the originally predicted star ratings whereas `neutral` only for the single intermediary one. Thus, it becomes apparent that the additional fine-tuning step with Moravian

data had a debiasing effect (see Figure 6.4). The confusion matrix of multilingual_sentiment_bert fine-tuned for Moravian sentiment also implies that the model has learned a slight `negative` bias. This bias is neither observable in the behavior of the fine-tuned model it is derived from nor the original base model when used in a zero-shot setting. However, it was observable when using bert, gbert, and bart as zero-shot sentiment classifiers. The other fine-tuned models, by contrast, rather learned a slight `positive` bias while sharing the fact that they rather confuse negative with positive samples than neutral with non-neutral ones (Brookshire and Reiter, 2024, 94). It is worth highlighting that a bias towards the label `positive` can also be attributed to the SVMs fitted on Moravian sources which implies that this could be related to the respective support in the training dataset (see Table 6.2).

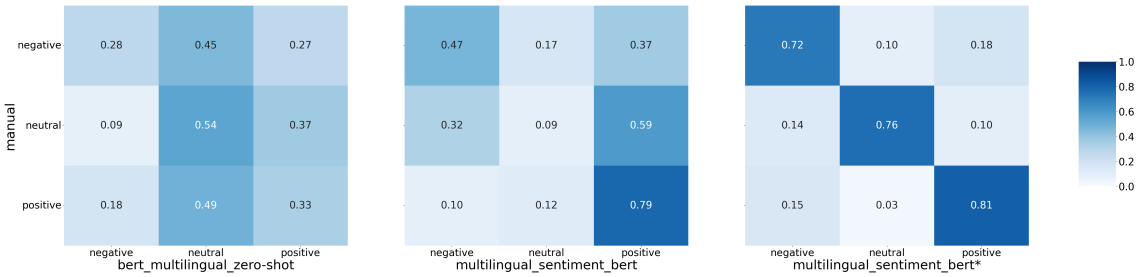

Figure 6.4: Influence of fine-tuning on the confusion matrices of the multilingual BERT models (systems fine-tuned on Moravian data marked with *)

Non-deterministic behavior of LLM classifiers and hallucination effects

One potential non-deterministic behavior of LLMs in general are so called *hallucination effects*. In the case of the zero-shot classification setting followed here, this means that the output is not one of the three sentiment labels. This was in fact not observed in any of the ten runs (= five per model). The only slight deviation from the prompted task was that the smaller LLM (= Gemma3_27B) added Unix line breaks (= `\n`) after the label in approximately 70 % of the cases no matter the run. As these errors are difficult to motivate but easy to postprocess, they are not counted as errors in Table 6.5 in order to not underrate the performance of the model.

The performance scores of Gemma3_27B still had a standard deviation of approximately 1 % because 16 sentences (= 4 % of the Moravian test dataset) were labeled differently between the five runs. Two sentences were even labeled with all three possible classes. The better performing Nemotron_Ultra_253B was more consistent since it only classified six test sentences differently between its five runs.

Close reading of misclassifications of the best model

Since the fine-tuned gbert model performed best, all further error analyses will focus on said system. In Brookshire and Reiter (2024, 94–95), we investigated the 84 cases where it predicted a class deviating from the manual annotation in a close reading step, and grouped them into four categories which account for 52 % of these errors (see Table 6.6). We disregarded the direction of the confusion (i.e. all `neutral` samples classified as `negative` were grouped together with `negative` samples classified as `neutral`).

As expected based on the confusion matrix, most confusions (= 44 %) arose when gbert needed to differentiate between the labels `negative` and `positive`. Yet, this confusion type is also the one which can be motivated the most because we found that the model struggled most with longer sentences. Example (6.4) is a typical one for this category as gbert attributed the label `positive` instead of `negative`. The reason for this kind of confusion may be due to the fact that the sentence includes positively connoted words like *liebe(n)* ‘dear’, *gefiel* ‘was pleased’, *sanft* ‘gently’, and *selig* ‘blessedly’ which are probably euphemistic or phrasal in nature (see Section 4.1.2). Since we decided to classify sentences by the way they are presented, this would in fact warrant the label

positive. However, we also decided to base the classification in mixed cases on the portrayal of the outcome (see Section 6.1) so that it still is a misclassification.

(6.4) *Am 11ten September 1778 muste sich meine liebe Frau an einer hitzigen Krankheit legen und am 29ten gefiel es dem lieben Heiland, sie sanft und selig zu vollenden, welches mir sehr schmerzlich war.*

('On the 11th of September 1778, my dear wife had to lay down with a severe illness and on the 29th, the dear Savior was pleased to complete her [life] gently and blessedly, which was very painful for me.')

Category	**Confusion**			**Total**
	negative ↕ neutral	negative ↕ positive	neutral ↕ positive	
long sentence	.07	***.20***	.02	.30
mixed sentiment	–	*.08*	.01	.10
multiple negations	–	*.05*	–	.05
short sentence	–	.01	*.07*	.08
other	**.14**	.10	***.24***	**.48**
Total	.21	*.44*	.35	1.00

Table 6.6: Relative distribution of misclassification categories of gbert fine-tuned for Moravian sentiment (highest amount per confusion in bold and per category in italics; corrigendum of Brookshire and Reiter 2024, 95)

A similar issue may have arisen in the case of the aforementioned Example (6.1) which is a more prototypical mixed sentence and was in fact also misclassified as positive instead of negative. We still decided to differentiate between these two categories despite a certain overlap as gbert was seemingly biased towards the label negative when classifying longer sentences but towards positive in the case of shorter truly mixed cases. More specifically, 44 % of the misclassified longer sentences (= 11 out of 25) were attributed the label negative but 75 % (= 6 out of 8) of the shorter ones positive. The latter tendency was even more pronounced for very *short* sentences, such as Example (6.5).

(6.5) *Ja, sagte Er, das thut tut Er an mir,*
('Yes, He said, He doth [sic!] does this to me,')

This sentence is rather neutral as their is no clear sentiment orientation due to the lack of context from the surrounding text passages. Yet, gbert classified it as positive reflecting its slight bias towards this label. Nevertheless, the reverse confusion with respect to a rather short sentence occurred one time, as well:

(6.6) *In dieser Hofnung lebte ich über ein halbes Jahr.*
('With this hoppe [sic!], I lived more than half a year.')

Example (6.6) is clearly positive but the model may have struggled with it due to the way *Hoffnung* 'hope' is written. More specifically, said word lacks one *f* (= the additional *p* for illustrative purposes in the translation) and this case of a historical graphematic variation may have impeded the identification of this decisive feature. This is due to the fact that the model splits the word into the two subword tokens Hof and ##nung which obfuscates the classification because *Hof* means 'court, yard, or farm' and may, thus, rather fit a neutral housing description.

Finally, there are a few instances which contain *multiple negations* and typically also past conditionals. Example (6.7) is a case of this kind and the sentence is arguably also difficult to label for human annotators.

(6.7) *hätte Er nicht Seine Hand über mich gehalten, wo wäre ich sündiger Mensch hingeraten.*
('if He had not held His hand over me, where would I, sinful human, have ended up going.')

In summary, the fine-tuned gbert model is not only the most performant system because it outperformed all other systems but also because quite a few of its misclassifications can be motivated. In this sense, some possible pitfalls for an application to new data were already identified. However, both the close reading step and the subsequent error categorization are not able to actually *explain* the actual classification behavior of said model. This is why the following section complements these error analyses with the help of explainability approaches.

6.3.3 Explainability analyses

The established explainability model SHAP (Lundberg and Lee, 2017) is accompanied by the Python module *shap* (version 0.43). It includes a default plotting function with which the decision-making that led to misclassifying the four previously discussed Examples (6.4), (6.5), (6.6) and (6.7) can be analyzed (see Figure 6.5).

Figure 6.5: Default SHAP explanation plots for selected misclassifications of gbert fine-tuned for Moravian sentiment (the predicted class is underlined at the top of each instance and the confidence highlighted in bold, negative SHAP values for the predicted class are colored blue and positive ones red, lighter colors correspond to SHAP values closer to 0)

In the case of Example (6.4) the SHAP values suggest that the fine-tuned gbert recognized all the positively connoted adjectives and adverbs as presumed. In fact, the word *sanft* 'gently' was actually the most decisive feature even though it is split into two subword tokens (= `san+ft`). This

shows that not all splits are equally severe due to the contextualized nature of BERT tokens. More specifically, the gbert token *san* followed by *##ft* occurs only in the case of other word forms of the related adjective *sanft* 'gentle' and noun *sanftmuth* 'gentleness' in the Moravian training dataset so that the model could learn their positivity.

SHAP also allows for a change of perspective by calculating the influence of subword tokens on a class different from the model attributed the highest confidence value to. The respective Figure 6.6 illustrates this in the case of Example (6.4) and the true label `negative` instead of the falsely predicted label `positive`. The respective shap values imply that gbert actually identified the darker note in the end of the sentence but it was outweighed by subword tokens leaning towards the false label `positive`. As the first half of the sentence was disregarded, it seems that the model may have learned the tie-breaking rule followed in this work.

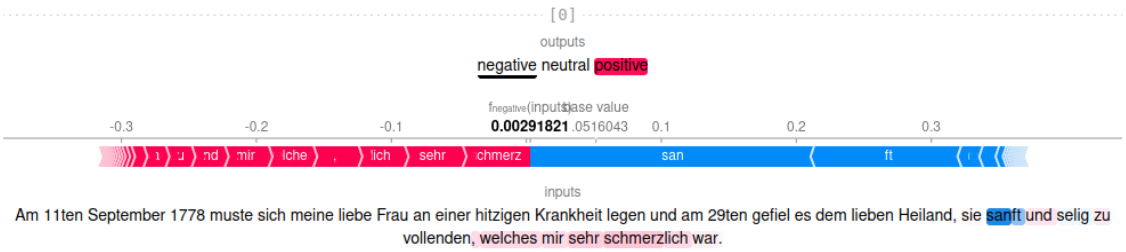

Figure 6.6: Default SHAP explanation plot of Example (6.4) with regard to the true label (underlined at the top, the confidence highlighted in bold, negative SHAP values for the predicted class are colored blue and positive ones red, lighter colors correspond to SHAP values closer to 0)

While changing the target label of shap values from the predicted to the true label shed more light on the behavior of gbert in this case, it does not add as much in the case of the other three examples. Nevertheless, a look at the impact of individual subword tokens on the predicted class helps to understand how their embeddings may have led to the respective misclassification. In the rather neutral Example (6.5), for instance, gbert assigned the positive class due to the three tokens *Ja* 'Yes', *tut* 'does', and *mir* '[to] me'. Among these three, the first is arguably the most plausible one for the confusion although contributing less than the other two. In this sense, said example hints at the problem that the model may have learned to associate actions directed towards the first person (narrator) with a positive polarity. This would be an unwanted bias if true.

The SHAP plot of the third example (6.6) supports the hypothesis from the previous section that the sentence is misclassified because it shares the token features *In dieser [...] lebte* 'lived in this' with neutral housing descriptions. It also shows that the first syllable of the word *Hof[f]nung* 'hope', which is *Hof* 'court, yard, or farm', was deemed more predictive than the second one, which is a typical ending of German deverbal nouns. Hence, the combination of these particularities seemingly lead to a split during tokenization that is more severe than the *san+ft* 'gentle' case.

The final example is worth mentioning because it is the only one of this small subsample where gbert attributed a rather low confidence value to the false label `negative`. This implies that the model had in fact issues with the negated past conditional so that the end of the sequence was considered for disambiguation reasons. More specifically, the model identified the tokenized phrase *sü+nd+iger Mensch* 'si+n+full human', which definitely has some negative connotation in textual sources in general, as being most predictive. However, the phrase is a rather typical collocation without any polar connotation in Moravian sources so that the debiasing through fine-tuning instances failed in this specific case. Yet, most impactful subword tokens identified with SHAP match expectations from the aforementioned close reading step. This suggests that (i) the model generates errors that are at least understandable and (ii) it "understood" the task at hand.

Leveraging Transformer-Explainability as a less resource-intensive alternative

Analyzing the four sentences with SHAP took approximately two minutes on a notebook CPU implying that this approach is difficult to follow in a real world scenario where a lot of samples

need to be explained simultaneously. Therefore, the results of a different explainability approach - namely the LRP implementation *Transformer-Explainability* by Chefer et al. (2021) - will be compared with the SHAP results in the following. The application on the four sentences discussed above took six seconds and the results are visualized in Figure 6.7.

text	label	negative	neutral	positive
Am 11ten September 1778 muste sich meine liebe Frau an einer hitzigen Krankheit legen und am 29ten gefiel es dem lieben Heiland, sie sanft und selig zu vollenden, welches mir sehr schmerzlich war.	positive	0.29%	0.08%	99.63%
Ja, sagte Er, das thut tut Er an mir,	positive	0.21%	0.95%	98.84%
In dieser Hofnung lebte ich über ein halbes Jahr.	neutral	0.10%	99.82%	0.08%
hätte Er nicht Seine Hand über mich gehalten, wo wäre ich sündiger Mensch hingeraten.	negative	58.16%	13.50%	28.34%

Figure 6.7: Transformer-Explainability based feature relevance table for selected misclassifications of gbert fine-tuned for Moravian sentiment (see Examples (6.4), (6.5), (6.6) and (6.7), lighter colors correspond to less decisive tokens)

When comparing Figure 6.7 with Figure 6.6 it becomes apparent that both approaches yield similar explanations. This is especially true for the first and third example as differences mainly concern the amount of impact individual features had on the prediction but not the overall implications regarding model behavior.

By contrast, in the case of Example (6.5), *mir* '[to] me' is deemed less decisive by Transformer-Explainability than by SHAP which would mean that the model did not learn this unwanted bias but instead based its misclassification more on the initial *Ja* 'Yes'. While this would still be a bias, it would be a more understandable one. It could even be exploited in a future debiasing effort during fine-tuning by adding more sentences with this structure to the other two classes.

Similarly, BERT-Explainability indicates that the model grasped the linguistically underlying conditional construction marked by the auxiliars *hätte* 'had' and *wäre* 'would' in the case of the final Example (6.7). This specific observation is not inferable from the SHAP values but could mean that gbert might have implicitly learned to associate this construction with the sentiment label `negative` during pre-training and/or fine-tuning. This is not unlikely since Manning et al. (2020) showed that learning linguistic structures is one of the emergent features of BERT models.

Scaling local explainability scores up to the corpus level

The explainability analyses described so far dealt with a small sample of individual model predictions because SHAP and BERT-Explainability are both local explainability models. However, it is also possible to scale these local explanations up to the corpus or specifically test dataset level in order to further approximate the general classification behavior. To this end, Brookshire and Reiter (2024, 95–96) compared the fine-tuned English BERT model with gbert. We first looked at the overall distribution of SHAP values per subword token and sentiment label (see Figure 6.8).

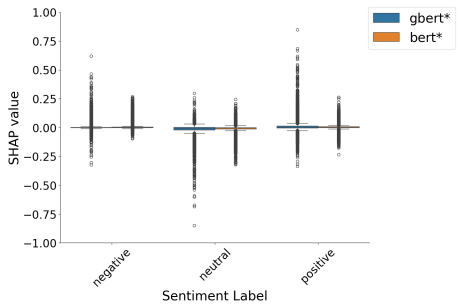

Figure 6.8: Distribution of mean SHAP values per subword token of bert/gbert fine-tuned for Moravian sentiment grouped by sentiment label (Brookshire and Reiter, 2024, 96)

The box plots illustrate that, no matter the sentiment label, only a few subword tokens were predictive. Instead, the vast majority of SHAP values are close to zero meaning that the corresponding

subword token did not impact the final label attribution on average. This is somewhat expectable given that only a few words actually bear sentiment information.

Interestingly, the distributions of mean SHAP values with regard to the label `neutral` are more negative than in the case of the two polar classes. This implies that both models attributed the former class more often when there are no subword tokens leaning towards one of the two sentiment poles. This is in line with the definition of a "truly neutral" class in Section 6.1 and hence a major learning success.

In a similar vein, it stands out that the range of mean SHAP values is larger in the case of gbert. This implies that it outperformed the English counterpart by an 18 % point F1 score margin because it has learned more discriminatory features during fine-tuning. In order to investigate this particularity even further, we zoomed in into the SHAP value distributions in a scalable reading sense by conducting an n-gram analysis. This analysis exploits the additive nature of SHAP values while accounting for the fact that BERT embeddings are contextualized. We found that the English model was more likely to follow random imbalances in the fine-tuning dataset. More specifically, it had issues with negators in negative and exclamation marks in positive samples which is in line with related work (Inácio et al., 2023). By contrast, gbert learned to identify typical German words bearing sentiment: With regard to the negative polarity, the following lemma forms occurred in subword bigrams with a mean SHAP value greater than .5: *Schmerz* 'pain', *verloren* 'lost', *Versuch[ung]* 'temptation', *Furcht* 'fear', *schwach* 'weak', and *con[fus]* 'con[fused]'. The most impactful positive words are *lieben* 'to love', *gut* 'good', *will+ig* 'will+ing', *Vergnügen* 'pleasure', *helfen* 'to help', *Dank* 'gratitude', and *Freude* 'joy' (Brookshire and Reiter, 2024, 96). These words yielded even higher SHAP values (above .66) in subword bigrams suggesting the slight positive bias mentioned earlier is based on the higher impact of positively connoted subword tokens.

Moreover, we assumed based on our findings that the perfomance difference between the two BERT models arose from differences between the respective tokenizers involved (Brookshire and Reiter, 2024, 98). More specifically, the English model splits more German words than gbert which increases the likelihood of severe splits of sentiment bearing words like in the aforementioned *Hof[f]nung* 'hope' case. This is due to the fact that the selection of subword tokens is based on their frequency in the pre-training datasets. Nevertheless, it is still striking that, despite this issue, the English model learned enough about sentiment in German Moravian sources during fine-tuning that it outperformed a modern-day German sentiment model and most lexical approaches by more than a 20 % point F1 score margin.

6.4 Summary

A ternary sentiment classification is a rather common text classification task. Yet, related work has shown that off-the-shelf systems are quite sensitive to the actual target language and domain (see Section 3.1.4). This is why a subset of the Moravian memoirs currently available in digital form were manually annotated first (see Section 6.1) in order to be able to evaluate said systems. These annotations comprised many non-neutral instances and hence underlined the relevance of *sentiment* as a concept with regard to the biographical records under investigation. In addition, a slight bias towards the label `positive` was observable which is another interesting finding.

Concerning automatic annotation approaches, the most promising ones discussed in related work were evaluated. These are (i) sentiment dictionaries, (ii) SVMs fitted to distributional features, (iii) fine-tuned transformer models, and (iv) zero-shot approaches (see Section 6.2). In line with related work, **dictionaries** performed rather poorly compared to other approaches as discussed in Section 6.3.1. This is noteworthy because they are preferred in digital humanities – most likely due to their transparent assignment of sentiment labels (Kim and Klinger, 2021, 11). When relating the performances to token and lemma coverage, it is worth adding that their was no clear correlation. Hence, a coverage analysis is only able to assess a priori whether a given dictionary will identify sentiment bearing words in general but not how well it will perform. Another notable finding is that the addition of contextual rules like in the case of GerVADER was beneficial. Still, better performances were reached with the two supervised learning approaches (= **SVM fitting**

and **transformer fine-tuning**). As a matter of fact, the latter was the most promising approach. However, it should also be mentioned that off-the-shelf transformer models, which have not been fine-tuned on data from the target domain before, performed significantly worse. This suggests that transformers have domain adaptation issues similar to the ones of dictionaries. Regarding **zero-shot approaches**, both evaluated LLMs performed surprisingly well although they showed a slight bias towards the label `neutral` (see Figure 6.2). They also showed less non-deterministic behavior than one might expect. This implies that they have "understood" the concept of *sentiment* to some extent – possibly because it is such a common task. However, the fine-tuned transformer model gbert still outperformed both LLMs despite (i) having less parameters by many orders of degree and (ii) requiring less computational resources.

A closer analysis of the classification errors of said most promising system (see Section 6.3.2) showed that it struggled most with (i) longer sentences, (ii) sentences with mixed polarities, and (iii) multiple negations (see Table 6.6). Especially the first two cases are ones where human annotators require an ambiguity resolution strategy which the model seemingly did not follow as consistently. In this sense, explainability analyses based on SHAP and Transformer-Explainability (see Section 6.3.3) underlined that gbert learned to distinguish sentiment bearing words from neutral ones since they yielded high impact/relevance values.

Chapter 7

Life stage classification

The second classification task of this work relates to the question of whether a person was a child or an adult when experiencing a given event described in a biography. This is called a **life stage classification**, here, and is – to the best of my knowledge – a novel task. Similar to the sentiment classification described in the previous chapter, it can be conducted by either manually or automatically annotating textual units. The operationalization used within this work is described in Section 7.1 for the former and in Section 7.2 for the latter case. Both sections focus on differences compared to the sentiment classification. Section 7.3 then deals with an evaluation of the selected automatic annotation approaches based on the manual annotations.

7.1 Manual annotation

In order to keep the manual annotations analyzable on their own and relatable to each other, the same subsample of Moravian memoirs (see Section 5) was chosen as for the ternary sentiment classification (see Section 6.1). Furthermore, the annotation of the current life stage of the person a given memoir is about was conducted sentence-wise, too. However, it is more difficult to infer the current life stage from a given sentence alone than a sentiment polarity. Thus, the sentences to be analyzed were kept in the original order and each text was annotated separately in order to grant access to more context. Still, the anonymization implemented by masking names was kept as is.

The actual annotation with life stages was two-fold and involved (i) the annotation of the current **age** of the biographical subject at the narrated time and (ii) a later **mapping to life stage labels**. This approach was deemed necessary because the selection of life stages is not undebatable (see Section 3.2.3). Hence, annotated age information is not only more intersubjective but also enables a different mapping in the future.

Age annotation

In a first step, all sentences which included explicit or implicit mentions of the subject's age were annotated accordingly. In order to account for implicit mentions based on dates or years, the annotator was given access to the year of birth of the biographical subject so that the current age could be calculated if necessary. This was helpful since many sentences of the annotated corpus include years like Example (7.1) or even exact dates like Example (7.2).

(7.1) *ao 1780 wurde ich zur Diaconissen eingesegnet.*
('in the yr 1780, I was consecrated as a deaconess.')

(7.2) *Den 2 Januar Anno 27 machten wir uns auf die reiße*
('On January 2, in the year 27, we set out on the journey')

In the first case, the biographical subject was born in 1727 so that one can conclude that she was 52 or 53 years old when becoming a deaconess. In the second one, the year is somewhat ambiguous due to being reduced to the decade, which is a common practice in these sources, but

most likely refers to the year 1727 given that the person was born in 1705. This means that the person was either 21 or 22 years old at the time she or he set out on a journey. Both sentences have in common that they are ambiguous with regard to the exact age when considering the respective year of birth. This kind of ambiguity was resolved by using the maximum age – i.e. 53 and 22 in these two examples.

It is worth mentioning that most sentences which include actual age information hint at the fact that year, dates, and even ages mentioned in the texts are not as exact as one could expect. This is shown in Example (7.3) where the specified age is qualified with the word *etwa* 'approximately' underlining that the author was not a hundred percent sure how old she or he was at the time.

(7.3) *Das war 1741. und ich war damals etwa 12. Jahre alt.*
('This was in 1741. and I was approximately 12. years old at that time.')

However, as a matter of fact, many sentences include neither explicit nor implicit information about the current age of the biographical subject. In these cases, further annotation iterations were necessary during which the most probable current age or age range was inferred from preceding and/or following sentences. This applies first and foremost to rather generic sentences, such as Example (7.4), but also to ones with relative temporal expressions like the phrase *ten years later*.

(7.4) *Er antwortete: Er solle sie von Ihm herzlich grüßen,*
('He answered: He should give her His warmest greetings,')

In some cases, it was impossible to identify the age of the biographical subject because the sentence to be annotated described a longer period of time like in the following Example (7.5). These cases were labeled with an age range and the disambiguation with respect to the task at hand was resolved during the subsequent mapping to life stages.

(7.5) *1759 kam ich zu Geschw. {NAME}s in Dienste, und bin 14 Jahr bey ihnen gewesen, bis sie Beyde zum Heiland gegangen waren.*
('In 1759, I came into the brethr. {NAME}'s service, and stayed with them for 14 years, until they Both went to the Savior.')

Since Example (7.5) is an excerpt from a memoir of a person born in 1714, it can be inferred that she or he was 45–59 years old when having worked with the other people mentioned and this age range was attributed in this annotation step.

Finally, there are sentences which do not deal with any biographical event at all – mostly because they are non-biographical remarks of the choir helper (see Section 4.1.2). These sentences were marked with the label `NONE` and a prototypical example is the following:

(7.6) *So weit ihr eigener Aufsatz.*
('So much for her own essay.')

Mapping ages to life stages

Having inferred the current age in the aforementioned way, the actual labels of this life stage classification task were attributed semi-automatically based on the age ranges listed in Table 7.1. This is why they are also the basis of the respective label definitions provided in Table A.2 in the appendix.

With the exception of the label `NONE` which accounts for non-biographic passages, the age ranges follow common subdivisions of the life course (see Section 3.2). More specifically, the first three stages were kept similar to the first four main ones of the MeSH thesaurus[23] whereas the latter three are the ones used by Haehner et al. (2024).

The first six years were not further subdivided as in the case of MeSH since the corpus comprised basically no references to *infancy* – i.e. individuals who are less than two years old –

[23] https://www.ncbi.nlm.nih.gov/mesh/68009273 (Last accessed on February 18, 2026)

other than birth descriptions. Hence, this differentiation does not appear to be as relevant in these sources as, for instance, the age of 13 which delimits the age groups *childhood* and *adolescence*. The latter was mentioned in many texts due to the occurrence of a religious rite of passage as shown in Example (7.7).

(7.7) *In meinem 13ten Jahr ging ich zum erstenmal mit zum heiligen Abendmahl*
('In my 13th year, I went to Holy Communion for the first time')

Label	Age	
	min	max
`early childhood`	0	5
`childhood`	6	12
`adolescence`	13	17
`early adulthood`	18	29
`adulthood`	30	59
`old-age`	60	–

Table 7.1: Age ranges of life stage labels

Since it was not only possible to annotate exact ages but also age ranges in in this pre-annotation stage, it was necessary to manually assign a single life stage label later if the annotated range spanned multiple ones. An example for this case of label ambiguity is the following:

(7.8) *in meinem 14ten Jahr lies mich mein Vater auch die Music lernen, zu meinem Schaden weil ich dadurch in böse Gesellschaften kam und 8 Jahre in dem Wirtshaus zum tantz gespielet habe.*
('In my 14th year, my father let me also learn music, which was to my detriment because it got me into bad company and I spent 8 years playing for dances in an inn.')

The subject of Example (7.8) was initially 14 years old and hence adolescent but a 22 year old adult by the time she or he left the inn, as indicated by the final phrase. Since the main biographical event occurred during adolescence, the sentence was labeled accordingly, and the same tie-breaking approach was used to convert all age ranges to a single age.

All annotated ages were automatically mapped to the respective label `early childhood`, `childhood`, `adolescence`, `early adulthood`, `adulthood`, or `old-age` based on the aforementioned minimum and maximum ages if not marked as `NONE`. The resulting per-class support is listed in Table 7.2.

Label	Dataset		
	Train	Test	Total
`early childhood`	88	19	107
`childhood`	149	39	188
`adolescence`	200	66	266
`early adulthood`	445	106	551
`adulthood`	**496**	**119**	**615**
`old-age`	243	63	306
`NONE`	147	30	177
Total	1768	442	2210

Table 7.2: Statistics of Moravian life stage classification datasets (highest support in bold)

As somewhat expectable, the label `adulthood` is the most common one and accounts for almost 28 % of the sentences. It is closely followed by `early adulthood` (= 25 %) which is more surprising because this stage accounts for only 12 years instead of 30. Hence, the memoirs

seem to focus on this stage and this finding can be related to prior research on Moravian memoirs (see Section 4.1.2). The same is true for the rather low amount of sentences dedicated to the early childhood which is even surpassed by choir helper remarks.

Table 7.2 also shows that the manually annotated sentences were split into 80 % training instances and 20 % held back for testing. This was implemented with the respective function of the Python module *datasets*, and involved a sentence-wise randomization like in the case of the sentiment classification.

7.2 Automatic annotation

For the sake of comparability, the same systems were chosen as in the case of the sentiment classification (see Section 6.2). However, since the life stage classification in the form used within this thesis is in fact a novel task, there are neither dictionaries nor other off-the-shelf systems readily available. Hence, the only systems to be evaluated in Section 7.3 are SVMs based on different distributional features (see Section 6.2.2), the transformer models gbert and bert_multilingual fine-tuned on Moravian data (see Section 6.2.3), and all six zero-shot models but the basic English BERT (see Section 6.2.4).

The latter model was not evaluated on this task for two reasons: First of all, it only led to a mediocre performance in the sentiment classification task no matter if used in a zero-shot setting or fine-tuned. Secondly, the explainability analyses conducted by Brookshire and Reiter (2024) suggest that the performance gap is due to its tokenization so that it is expected to perform worse in this second task, as well.

In the case of the SVMs and the selected base transformer models, the manual life stage annotations presented in Section 7.1 were used for training. In the case of the evaluated LLMs, the basic prompt template shown in (6.2) was adapted by replacing the abstract task name `[C]` with *life stage* and constraining the answer space `[Y]` to the seven life stage labels in the following way:

(7.9) *Please perform a life stage classification task. Given the sentence, assign a life stage label from ["NONE", "adolescence", "adulthood", "childhood", "early adulthood", "early childhood", "old-age"]. Return label only without any other text. Sentence: [X] ... Label:*

Although the selection of systems suggests a high degree of similarity to the other two tasks conducted within this thesis, it needs to be stressed that the automatic annotations differ from the manual operationalization only in the case of this specific task. More specifically, the year of birth and surrounding sentences where provided in the case of the manual annotations in order to make them as intersubjective as possible. However, a similar practice would have led to less randomized and potentially also smaller training and test datasets. This has implications for systems, such as transformer models, which can later be adapted to other tasks. One potential issue is that longer excerpts of the original biographical records may leak when later applied for text generation tasks. A different problem is that additional context may come at the cost of (potentially) adding more ambiguities because more sentences may include more temporal expressions while also leading to a slightly different input format compared to the other tasks. This does not mean that future research could not find ways to close the gap between the manual and automatic annotations presented here. However, in order to gain a deeper understanding of the framing of life stages in biographical sources by relating them to the other two tasks, only considering one sentence at a time is seen as some kind of a compromise for the current work.

7.3 Evaluation

Similar to Section 6.3 which dealt with the evaluation of automatic approaches to a sentiment classification, this one looks at the performance of automatic life stage classifiers. To this end, their annotations are compared with the aforementioned manual ones in Section 7.3.1. Some notable classification errors of the most promising systems are then further analyzed in Section 7.3.2.

Section 7.3.3 builds up on this error analysis by illustrating how the classification behavior of the most performant system, which is once again the fine-tuned gbert model, can be explained using Transformer-Explainability.

7.3.1 Performance comparisons

The evaluation of the aforementioned automatic annotation approaches is based on weighted precision, recall and F1 score as calculated by the Python module *sklearn* with no preprocessing. The results are summarized in Table 7.3 which adds a random and majority baseline with the latter one always attributing the label `adulthood`. The results of the 36 evaluated SVM parameter settings are reduced to the best one per analyzer in the table but fully shown in Figure 7.1. In the case of the evaluated LLMs, the mean and standard deviation of five different runs on five different days are reported in order to account for their non-determinism.

System	Type	Precision	Recall	F1 score
Random	baseline	.19	.15	.16
Majority	baseline	.07	.27	.11
SVM_char2-3_TF-IDF*	SVM*	**.56**	.44	.43
SVM_word1-3_count*	SVM*	.47	.45	.43
bert_multilingual*	BERT*	.36	.43	.37
gbert*	BERT*	.52	**.50**	**.49**
bert_multilingual_110M	zero-shot (BERT)	.17	.10	.05
gbert_110M	zero-shot (BERT)	.17	.15	.13
bart-nli_406M	zero-shot (NLI)	.15	.14	.11
mDeBERTa-nli_279M	zero-shot (NLI)	.25	.24	.23
Gemma3_27B	zero-shot (LLM)	$.36_{\pm.02}$	$.29_{\pm.01}$	$.25_{\pm.01}$
Nemotron_Ultra_253B	zero-shot (LLM)	$.41_{\pm.02}$	$.17_{\pm.02}$	$.16_{\pm.02}$

Table 7.3: Life stage classification results (best value per score in bold, systems fine-tuned/trained on Moravian data marked with *, only the best results per analyzer are listed in the case of SVMs)

Most results fit the assumption that this task is rather difficult because almost all performance scores are quite low. More specifically, both basic BERT models as well as BART performed worse than the random baseline when used in a zero-shot setting. Interestingly, the LLM with the highest amount of parameters (= Nemotron_Ultra) also shows a similar low recall and F1 score although its precision is with .41 higher and only outperformed by SVMs and the fine-tuned gbert model. This tendency of a higher precision than recall is shared by all systems but the fine-tuned bert_multilingual. The German counterpart of the latter (= gbert) is the most performant system.

As a whole, the performance scores led to a clear ranking between the different system types: Systems trained in a supervised fashion performed better with fine-tuned BERT models yielding scores in the range .36 to .52 (mean .45) and SVMs in the range .21 to .56 (mean .39) whereas zero-shot approaches yielded the poorest scores ranging from .05 to .41 (mean .20).

Performance of SVMs

Regarding the different evaluated SVM settings, it is worth pointing out that the performance of the word analyzer stayed quite consistent when adding bigrams and/or trigrams to unigrams but dropped when removing the latter (see Figure 7.1). This uniformity is noteworthy because the three n-gram ranges lead to different dimensionalities (see Table 6.4).

The other evaluated parameter (= vectorizer weighting) did not affect the dimension and appears to only have a minor impact in this setting. This is different in the case of the second evaluated analyzer (= a character-based vectorization) where a TF-IDF weighting underperformed compared to the other two simpler vectorizations when combining uni- and bigrams but lead to a slight

performance gain when combining bi- and trigrams. Regarding the n-gram ranges, a character-based vectorization profited from a higher n than the word-based one with performance drops only observable when relying on unigrams and/or bigrams alone.

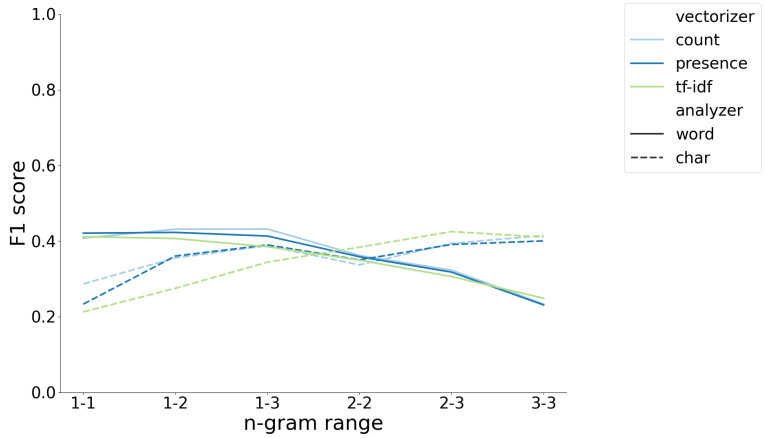

Figure 7.1: Influence of different parameter settings on SVM perfomances

In summary, these mixed results underline the usefulness of a grid search – even more so because it does not require many computational resources in the case of SVMs trained on distributional features of a dataset with a size similar to the one of this work.

Performance of fine-tuned transformer models

Both evaluated transformer models outperformed zero-shot settings when fine-tuned with data from the target domain by a 19–36 % point margin (see Table 7.3). The lowest performance gain relates to the precision of the multilingual model where no difference was observable in the zero-shot setting. Still, this transfer learning approach appears to be worth the effort – even more so when considering that its German counterpart yielded the second best precision as well as the best recall and f1 score among all evaluated systems when fine-tuned.

Performance of zero-shot approaches

In the case of the other two evaluated zero-shot approaches which are prompt-based LLMs on the one hand and an exploitation of the NLI task on the other, it is striking that models with less parameters outperformed ones with more. The only exception is the comparably high precision of Nemotron_Ultra. Both observations also apply to the slightly higher standard deviation of Nemotron_Ultra compared to the one of Gemma3. However, most scores are much lower than the ones of the supervised learning settings.

7.3.2 Error analyses

While performance scores offer a rough quality assessment, the confusion matrices depicted in Figure 7.2 also hint at typical error types. The figure deals with the best SVM setting (= TF-IDF weighted character bi- and trigrams), fine-tuned transformer model (= gbert), and best LLM run (= the third of Gemma3).

As expected based on a .25 F1 score which is only 9 % points above random, the most performant zero-shot approach (= Gemma3) confused many labels and basically even failed to handle sentences dealing with people in *early adulthood*. However, it is at least competitive to the other two better performing approaches when predicting the labels `childhood`, `adulthood`, and `old_age`. Part of the reason seems to be that it favored the former two as well as `NONE` in general. This latter bias relates to the recall issue which was less prominent in the case of the best SVM setting as well as the fine-tuned gbert model.

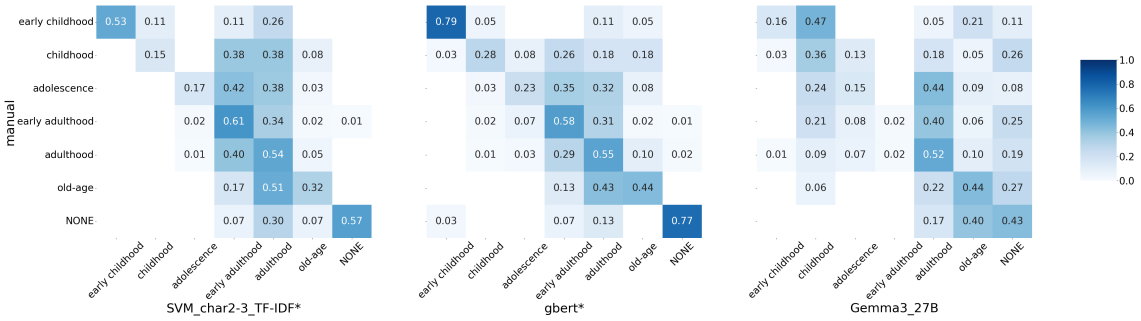

Figure 7.2: Confusion matrices of best performing life stage classifiers per type (systems fine-tuned/trained on Moravian data marked with *)

It should be noted that both supervised learning approaches struggled most with the labels `childhood` and `adolescence` which have a rather low support in the train dataset. Instead, they showed a slight bias towards the labels `(early) adulthood` which has the highest support (see Table 7.2). This bias is more pronounced in the case of the SVM classifier. However, the labels `early childhood` and `NONE` are in fact the two with the lowest support and yet the SVM and even more so the fine-tuned gbert confused respective sentences rather seldom with other classes. Hence, support in the training dataset does not seem to be the sole predictor as it rather appears to be the case that some classes are more difficult to distinguish from each other than others. More specifically, this seems to affect subsequent life stages. This is why off-by-one errors are ignored in the following to get a deeper understanding of possible further error sources.

Off-by-one errors

In order to assess off-by-one performances, all predicted labels but `NONE` were also counted as "correct" whenever a life stage preceding or following the true one was predicted (e.g. `old_age` instead of `adulthood`). This even more coarse-grained evaluation yielded better results for all systems as illustrated by a comparison of Table 7.3 and Table 7.4.

System	Type	Precision	Recall	F1 score
Random	baseline	.48	.38	.42
Majority	baseline	.50	.65	.55
SVM_char2-3_TF-IDF*	SVM*	**.83**	.79	.76
SVM_word1-3_count*	SVM*	.75	.75	.74
bert_multilingual*	BERT*	.75	.80	.76
gbert*	BERT*	**.83**	**.82**	**.81**
bert_multilingual_110M	zero-shot (BERT)	.70	.29	.26
gbert_110M	zero-shot (BERT)	.60	.36	.42
bart-nli_406M	zero-shot (NLI)	.57	.40	.42
mDeBERTa-nli_279M	zero-shot (NLI)	.61	.58	.56
Gemma3_27B	zero-shot (LLM)	$.70_{\pm.01}$	$.53_{\pm.01}$	$.57_{\pm.01}$
Nemotron_Ultra_253B	zero-shot (LLM)	$.78_{\pm.00}$	$.25_{\pm.04}$	$.29_{\pm.05}$

Table 7.4: Life stage classification off-by-one results (best value per score in bold, systems fine-tuned/trained on Moravian data marked with *, only the best results per analyzer are listed in the case of SVMs)

Whereas the recall issue basically prevails, different systems profited differently from this task simplification. Most prominently, the zero-shot classifiers bert_multilingual and Nemotron_Ultra profited the least. This implies that they did not only fail to identify relations between sentences and labels, but also the inherent ones between the labels of this task.

Non-deterministic behavior of LLM classifiers and hallucination effects

Even though both LLM classifiers performed poorly on this specific task, it is worth mentioning that their non-deterministic nature only led to minor standard deviations of 1–2 % points from their mean performance scores. This is a little higher but still similar to the ones observed when applied to sentiment (see Section 6.3.2). However, in the case of Nemotron_Ultra, the standard deviation increased to .04 (recall) and .05 (F1 score) when neglecting off-by-one errors.

In spite of this difference, it is striking that the annotations of Nemotron_Ultra varied in less cases (= 78 or 18 % of the test sentences) between the five individual runs than the ones of Gemma3 (= 108 or 24 %). In this sense, the latter model was less deterministic as it attributed different labels between runs to almost every fourth instance but was still able to yield a better recall per run than the bigger model. It should also be noted in this context that line breaks were cleaned before the performance evaluation in order to not underrepresent the capacity of Gemma3 as was done for the sentiment classification (see Section 6.3.2).

A "true" hallucination effect was observable for this life stage classification task as the larger model Nemotron_Ultra introduced the **German** label `Jugend` 'youth' in the first two runs when annotating Example (7.10). In the final three runs, it attributed the label `adulthood` instead which is an off-by-one error as the respective person was 71–78 years old at the time she or he expressed the request described in the sentence to be annotated.

(7.10) *Die Ursach, warum ich ersuche, unter den Brüdern ein Ruheplätzgen für meine zurückgelaßene Hütte zu bekommen, ist, weil mich von meiner ersten Erweckung in meiner Jugend, an das große Unterscheidungs- Kenn-Zeichen derselben, unsers lieben Heilands verdienstliches Leiden und Tod unter allen Schwierigkeiten stets so vest an sie gebunden hat, daß ich mich bis auf diesen Tag als das unwürdigste Mit-Glied zu der Brüder-Gemeine bekenne*

('The reas'n why I request a resting place for my abandoned hut among the Brethren is because, from my first awakening with its great distinguishing- hall-mark [sic!] in my youth on, our dear Savior's meritorious suffering and death has always bound me so firmly to them under all difficulties that I confess myself as the most unworthy mem-ber [sic!] of the Brethren community to this day')

Close reading of misclassifications of the best model

As inferable from the performance scores, the fine-tuned gbert showed the least amount of confusions. Yet, it still confused quite a few life stages although the most common confusions are off-by-one errors (see Table 7.5).

Confusion	Off-by-one	Occurences
`early adulthood` ↔ `adulthood`	yes	68
`adulthood` ↔ `old-age`	yes	39
`adolescence` ↔ `early adulthood`	yes	30
`adolescence` ↔ `adulthood`	no	24
`childhood` ↔ `early adulthood`	no	12
`early adulthood` ↔ `old-age`	no	10
`childhood` ↔ `adulthood`	no	8
`childhood` ↔ `old-age`	no	7
`adulthood` ↔ `NONE`	no	6
`childhood` ↔ `adolescence`	yes	5
`adolescence` ↔ `old-age`	no	5
`early adulthood` ↔ `NONE`	no	3
`early childhood` ↔ `childhood`	yes	2
`early childhood` ↔ `adulthood`	no	2

Table 7.5: Confusions of gbert fine-tuned for Moravian life stages that occurred at least twice

When disregarding off-by-one errors, it seems as if gbert classified many childhood experiences as having occurred at some later life stage. A close reading of the respective 27 sentences showed that in fact only two sentences are obvious errors because they include the age of the child. The other misclassifications of this type are generic ones which is why they are quite difficult if not impossible to associate with one life stage or another without context. One such example is Example (6.6), which was also an error in the sentiment classification task.

As gbert attributed the two labels `early adulthood` and `adulthood` to most of these sentences, it seems as if gbert learned to use them as some kind of default for ambiguous cases. This is an understandable choice given that both life stages do not only account for 53 % of the training instances (see Table 7.2) but also the majority of the life span of most human beings.

In a similar vein, it is understandable to a certain degree that gbert appears to have learned to associate more generic descriptions of (severe) health issues with the label `old-age` in five cases all of which are similar to the following Example (7.11):

(7.11) *Dazu noch das kam, daß eins von seinen Beinen zu schwöllen anfing.*
('On top of that, one of his legs started to swell.')

In summary, this means that the fine-tuned gbert definitely led to a few misclassifications which are wrong in a more or less obvious way. However, most misclassifications can be either motivated by being off-by-one errors, by involving rather generic descriptions, or by dealing with a topic a human might also associate with another life stage without context. While these "explanations" seem plausible, the letter two as well as the more general classification behavior are investigated further with a true explainability approach in the following section.

7.3.3 Explainability analyses

As a first step towards explaining the two aforementioned exemplary misclassifications from internal states of gbert alone, it is insightful to look at the distribution of output probabilities listed in Figure 7.3. They show that the correct class was the second most likely one in both cases which could imply that the model found some hints for the respective correct class. However, the difference to the wrongly attributed class is still notable. Therefore, it is even more interesting to look at the subword tokens contributing most to the actual output as inferred with the LRP-based Transformer-Explainability approach.

text	label	early childhood	childhood	adolescence	early adulthood	adulthood	old-age	NONE
In dieser Hofnung lebte ich über ein halbes Jahr.	early adulthood	10.34%	22.69%	3.19%	61.15%	1.27%	1.27%	0.09%
Dazu noch das kam, daß eins von seinen Beinen zu schwöllen anfing.	old-age	4.11%	17.17%	15.22%	16.88%	6.29%	39.88%	0.45%

Figure 7.3: Transformer-Explainability based feature relevance table for selected misclassifications of gbert fine-tuned for Moravian life stages (see Examples (6.6) and (7.11), lighter colors correspond to less decisive tokens)

In the first case, gbert appears to have based its decision-making on the phrase *in dieser Hof[f]nung lebte ich* 'I lived in this ho[p]e' with a focus on the verb. The verb appeared only eleven times in the training dataset and even though three of the respective sentences are about early adulthood the same is true for the residual class. So, there is no obvious imbalance in the train dataset and the label `early adulthood` was probably really chosen as some kind of default in this specific case.

Regarding the second exemplary misclassification, it is noteworthy that the explainability model detected some focus on the noun phrase *seinen Beinen* 'his legs'. As body parts are likely to appear in the context of health issues, this may support the assumption from the previous section that the model associated health issues with an old age. Yet, the main cue towards health issues *schw[e]llen* 'swell' seems to be less relevant for the actual decision-making – possibly because this verb is split during tokenization.

Somewhat surprisingly, it looks like the model also considered the final dot when classifying. This may hint at an unwanted bias but does not seem to correlate with an imbalance in the train dataset as the label distribution of a final dot is identical to the overall distribution listed in Table 7.2.

Scaling local explainability scores up to the corpus level

The two exemplary local explainability analyses show that Transformer-Explainability assigns relevance scores to individual subword tokens in a similar way to SHAP although there are some differences between both approaches (see Section 2.3.4): Most notably, Transformer-Explainability is computationally less demanding but does not account for negative impacts. While the former peculiarity is an advantage no matter the classification task at hand, the latter one is problematic when trying to explain a sentiment classifier with its `neutral` class since it is defined ex negativo (see Section 6.3.3). However, this disadvantage is less pronounced in the case of a life stage classifier because it distinguishes between more classes which are very likely characterized by relevant tokens. This may even apply to the `NONE` class used for non-biographical remarks as they may be phrased in a way which is quite distinct from the one used in the context of the other classes. Thus, the relevance scores inferred by Transformer-Explainability are expected to be quite insightful not only when used to look at individual misclassifications locally but also when scaled up to the whole test dataset in order to approximate a global explainability of the decision-making. To this end, the distributions of relevance scores per life stage label were assessed.

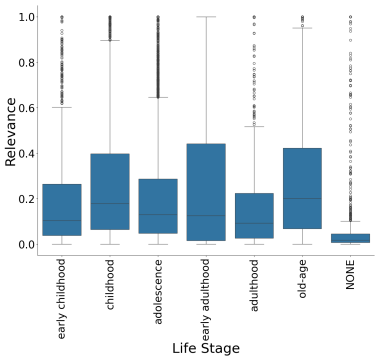

Figure 7.4: Distribution of relevance scores per class for gbert fine-tuned for Moravian life stages applied to the Moravian test dataset

Figure 7.4 shows that Transformer-Explainability attributed the maximum relevance score to at least one token no matter the class. These highly relevant tokens are, however, mostly outliers which is somewhat expectable given that most subword tokens should be generic enough to appear in descriptions of any events happening in any life stage.

Likely for similar reasons, the mean per-class relevance scores are rather low even though one can observe interesting inter-class differences: For instance, the interquartile range of relevance scores for assigning the labels `adolescence`, `early adulthood`, and `old-age` are larger than the others, while the residual class `NONE` has the smallest and by far lowest one. Yet, the fine-tuned gbert is still able to discriminate rather well between instances of this `NONE` class and ones of the actual life stage classes given the confusion matrix depicted in Figure 7.2. However, the highest mean relevance score is attributed to subword tokens found in sentences describing events during old age and the second highest to ones about childhood experiences. Therefore, it seems as if the model learned to identify more discriminating textual features between these two classes than inferable from the confusion matrices alone.

In order to further investigate the decision-making process, the most relevant individual subword token features were assessed, and grouped by lexical fields (see Table 7.6). In this context, it was rather difficult to identify (interpretable) fields which gbert could exploit to distinguish, for instance, between descriptions of early childhood and old age. More specifically, the subword tokens deemed most relevant for the latter comprise mainly pronouns and otherwise ambiguous words instead of ones which are connected to this life stage in an obvious way.

Similar issues arose in the case of the other classes. This is rather surprising given the fact that the model was able to learn some notable textual features of life stages based on its (off-by-one) performances. Hence, the remainder of this section will try to connect some of the inferred lexical fields with the respective life stages:

Starting with the very first one, it seems like the model may have associated health-related vocabulary as well as the body part *Beinen* 'legs' not only with the class `old-age` in sentences like Example (7.11) but also with the class `early childhood` (see Table 7.6). This is somewhat unexpected but might relate to the fact that child diseases were common in the 18th century and could hence appear frequently in these sources. This is also part of the reason why a few memoirs in the corpus are about Moravians who already died in this early stage of their life.

While there are no health-related tokens in the list of the subsequent `childhood` stage, there is another body part – namely *Kopf* 'head' – which could hint at a similar connection. In addition, there are movement words, such as `verließ` 'left' or ones like `aufgenommen` 'accepted'. The former is commonly used in euphemistic ways and the latter in religious contexts so that they are difficult to categorize.

Words describing human relationships and even more so romantic ones like `Heirat` 'wedding' seem to be misplaced when looking for relevant textual features of childhood experiences. Yet, they may play a role here because Moravians tended to interweave information about parents with them. This genre-specific feature complicates the classification task even more but may hint at another reason for the observed default status of the classes `(early) adulthood` besides the already mentioned considerations regarding age spans and support in the dataset used for training.

Content-wise, it is worth mentioning that the classes `childhood` and `adolescence` are the ones where *locations* and *movements* appear to be more relevant than in the case of the other life stages. This is somewhat unexpected given the fact that Moravians also tended to move around for work. So, maybe the model found different phrasings when describing different life stages which could be worth analyzing further. The same is true for the similar distribution of *religious* vocabulary which is additionally deemed relevant for the class `old-age`.

Finally, it should be highlighted that the two classes `early adulthood` and `adulthood`, which have the highest support in the training data, also share a rather low amount of most relevant tokens and a slight focus on kinship terms. This can be motivated with the higher probability of family-related events, such as the birth and growing of children, occurring during these two life stages. In addition, the high degree of similarity between the two per-class relevance lists may explain the high number of confusions between these two classes.

7.4 Summary

A life stage classification is a novel task introduced in this thesis which accounts for genre-specific features of biographical records. To this end, the corpus described in Chapter 5 was first manually annotated sentence-wise with subject ages. This information was either extracted from explicit references or inferred from temporal dependencies between contextual phrases (see Section 7.1). Afterwards, it was mapped to life stage categories proposed in related work (see Section 3.2.1).

Content-wise, the label distribution showed that the annotated Moravian memoirs focus on (early) adulthood experiences (see Table 7.2). Moreover, it is noteworthy that 8 % of the sentences were in fact not assigned any life stage label at all because most of them were non-biographic remarks of the choir helper(s) who composed the final memoirs from (auto)biographical notes.

In order to be able to scale such analyses up to more texts, the manual annotations were also used to train SVMs with distributional features and to fine-tune transformer models. Furthermore, a few zero-shot approaches were evaluated. All systems yielded rather low (raw) performance scores (see Section 7.3.1). This implies that the task is considerably more difficult than, for instance, a ternary sentiment classification.

Regarding performance differences, it is worth mentioning that **zero-shot classifiers** in general and the two evaluated LLMs in particular yielded very poor results no matter the amount of parameters involved. Moreover, the LLMs showed notable non-deterministic behavior as they

classified 18–24 % of the test instances differently between different runs. The larger LLM Nemotron_Ultra even introduced the label `Jugend` 'youth' which is German unlike the others. Hence, LLM classifiers are by far less promising for this task than in the case of the sentiment classification discussed in the previous chapter.

The best **SVM** setting as well as **fine-tuned transformers**, which are both supervised learning approaches, by contrast, appear to have learned how to discriminate between different life stages – at least to some degree (see Figure 7.2). The main issue these models had was the discrimination between consecutive stages. This can be explained with the fact that subdividing life courses is per se difficult, for instance, because it does not factor in individual biographical differences. Considering this and similar backdrops, the performances of the evaluated SVMs and fine-tuned transformer models are in fact surprisingly high. This is even more true when excluding off-by-one errors (see Table 7.4).

In the case of the fine-tuned gbert model, which performed better than the rest, a close inspection of typical errors hinted at potential biases (see Section 7.3.2). One such bias is that the classes `early adulthood` and `adulthood` were seemingly regarded as some kind of default choices. In addition, health issues were associated with the class `old age`. Subsequent (global) explainability attempts were performed by identifying most relevant subword tokens and grouping them by lexical fields (see Section 7.3.3). This hinted at the possibility that the model deemed health issues even more relevant for the class `early childhood`. While all these biases are less straight-forward than the ones found in the sentiment classification task, they are at least reasonable given the inherent task difficulty. In this sense, they underline that raw performance scores alone may not always be enough to validate whether a given classifier fits the expectations or not.

Label	Subword Tokens	Lexical Field
early childhood	Wir, unseren, seinen, wir, sein, Er, er, ihn, ihm	pronoun
	##la, ##mm, gehei, antwortete, Mann, Sachen, heim, Ruhe, ##dg, einzurichten, erh, ##samm	*ambiguous*
	Alter, Uhr	age/time
	Atem, krank, Kräfte, Leben	health
	Beinen	body part
childhood	##42, Frühjahr, ##43	date/number
	##än, sind, sah, kennen, hei, folgendes, erlaub, ehrlich, Camer, Erlaubnis, ##aden, ##r, ##ach, waren, Gedanken, allen, ##mme, con, ##bir, nimm, ##serie, trat, ##uß, getr, lebte, lag, ##rate, ##uff, geriet, ##iel, ##zl, wünschte, get, ##aßen, ##mün	*ambiguous*
	Kopf	body part
	Heirat, Ehe, Freund, Freundschaft	relationship
	Gang, verließ, besuchten, gekommen, flüchte, gelangte	movement
	Eingang, Holland, zwischen, Farm, haus	location
	Got, Aufnahme, aufgenommen, glaubte	religion
adolescence	zugesprochen, erlebte, Nil, Comp, Thron, Auflösung, einmal, ##ritten, ##lin, genießen, getan, suchte, ber, ##agn, Anges, ##ading, fanden, Seht, angetreten, ##gem, ##con, Von, ##reis, Dia, gewährt, ##denken, Mus, zogen	*ambiguous*
	October, December, gegenwärtig	date/time
	##häus, England, Plant, York, Ufer	location
	Krieg, Hauptmann	military
	Bitte, offenbart, Chore, Seel, Bestimmung	religion
	##reisen, Bewegung, verließen, ##fahrt, Reise, Besuch	movement
	##macher	occupation
early adulthood	##ern, ##lt	*ambiguous*
	Kindern, Vater, Mutter, Kinder	kinship
	8, Aug, 4, April	date/number
	geboren	birth
	sie	pronoun
adulthood	8	date/number
	Eltern, Vater, Mutter	kinship
	Schule	location
	unser	pronoun
	gebracht, schickte, geschl	*ambiguous*
old-age	Wir, dich, deinem, unser, ihrer, Unser, du, sie, Sie	pronoun
	Von, ##iger, weit, schreibt, Nach, rufen, ##ia, wäre, ##nehmen, O, ##bew, Personal, lieb, ##ster, ##laß, folgender, sagen	*ambiguous*
	Lamm, Schöpf	religion
NONE	mein, Ich	pronoun

Table 7.6: Subword tokens with a minimum relevance of .75 for gbert fine-tuned for Moravian life stages

Chapter 8

Life event domain classification

In a third and final text classification task, the subsample of Moravian memoirs described in Chapter 5 was annotated with ten different life event domain labels. The selection of labels and subsequent annotation procedure is described in Section 8.1. Afterwards, Section 8.2 deals with the selection of automatic classification approaches whereas Section 8.3 presents performance, error, and explainability analyses. The following Chapter 9 then shows how both the manual and automatic life event domain annotations can be related to ones of the other two tasks in order to identify patterns in biographical sources.

8.1 Manual annotation

A life event domain classification was operationalized by slightly adapting 13 main categories Haehner et al. (2024) inferred from open responses to contemporary household surveys (see Section 3.3.1). This novel operationalization was preferred over existing event type classification schemes in order to account for needs from both biographical and historical research (see Section 3.2.3). Ten categories from the original work (= `collective`, `education`, `family`, `health`, `housing`, `leisure`, `money`, `other_social`, `romantic`, and `work`) were unchanged. In practice, this means that the original definitions previously shown as Examples (3.11) to (3.20) were reused (see Table A.3 in the appendix for the exact phrasing used here).

The original category `other` was split into the original subcategories `religious` and `legal` as defined in Examples (3.24) and (3.25). In the case of the first former subcategory, this was done in order to account for the religious context of the texts to be analyzed whereas the second one was mainly lifted to a main category by analogy. *Religious* events are also noteworthy because they may include imagined events, such as dreams and visions, which are also relevant for Moravian sources (Faull, 2021, 217).

The two residual categories `Not clear: poor information` for truly ambiguous cases and `Not clear: more than one domain` for mixed ones were combined to the new `no_or_other` class. This alteration of the original categorization increases the similarity to the tasks described in the previous two chapters as they also only encompassed a single residual class. The new class was described in the annotation guideline of this work in the following way:

(8.1) *all cases not covered with one of the other labels.*

Finally, the two classes `birth` and `death` were added as both are essential events when categorizing biographical records dealing with the whole life span. In the original study, by contrast, they were irrelevant because people were asked to describe the most important event which happened in the previous year. In this context, it is not expected that someone answers with phrases, such as *I was born* or *I died*. Yet, both events were mentioned in the definition of the class `family` defined in Example (3.13), and hinted at in the case of the classes `romantic` and `other social`. The definitions of these two classes are as follows (see also Table A.3):

(8.2) *the birth/death of the biographical subject.*

Having defined all 15 initial or "raw" labels, the manual annotation itself was conducted in a similar way to the one described in Section 6.1. In practice, this means that the 2210 sentences subsampled from Moravian memoirs (see Section 5) were annotated in isolation unlike in the case of the life stage classification (see Section 7.1). For disambiguation purposes, only the last event was considered whenever multiple appeared in a given sentence. One example from the Moravian dataset per life event domain is listed in Table 8.1. Please not that all translations again try to preserve as many instances of non-normative spellings and punctuation as possible.

Label	Text	Translation
`birth`	Ich bin den 8. Oktober 1717 zu Balkheim im Fürstentum Oettingen geboren.	I was born the October 8th 1717 in Balkheim in the Principality of Oettingen.
`collective`	so bald wir wieder zu hause gekommen, hörten wir /// daß Indianer Krieg werden würde,	as soon as returning to our home, we heard /// that there would be Indian war,
`death`	Am 13. abends verschied er sehr sanfte u. sein Ende zeigte, daß er geglaubt.	On the evening of the 13th, he passed away very peacefully a. his end showed he had believed.
`education`	Meine Lehrzeit war 5 Jahr, und ein Viertel Jahr bin ich noch nach Endigung derselben, da geblieben.	My apprenticeship lasted 5 year, and I stayed there even after its ending for a quarter of a year.
`family`	Ihre Ehe war mit 4 Kindern gesegnet.	Her marriage was blessed with 4 children.
`health`	da lag ich 3 Wochen sehr kranck,	I lay down very sic for 3 weeks,
`housing`	1760. zog sie nach Naz.	1760. she moved to Naz.
`legal`	–	–
`leisure`	Indessen besuchte ich bald den Herrnhag.	Meanwhile, I soon visited the Herrnhag.
`money`	Im Jahr 1733 kaufte mein Vater 100 Acker Land für 15 10/-.	In 1733, my father bought 100 acres of land for 15 10/-.
`other_social`	Mit Vergnügen trafen wir hier die Geschwister am Bau ihrer Winterhäuse arbeiten;	We were pleased to meet the siblings here building their winter homes;
`religious`	ao. 1762 wurde ich zur Acoluthie angenommen.	ao. 1762, I was accepted as acolyte.
`romantic`	Die Trauung erfolgte am 16. Aug. des gedachten Jahres 1770.	The wedding was held on Aug. 16 of the commemorated year 1770.
`work`	Hier stund er eine Zeitlang der Schneiderey als Meister vor.	Here, he headed as a master tailor for a While.
`no_or_other`	Mein Nahme war {NAME}.	My naame was {NAME}.

Table 8.1: Examples from the manually annotated Moravian dataset per (raw) event domain class

Reducing the number of life event domain classes

Table 8.1 lacks an example of a *legal* event because actually none of the 2210 sentences were labeled accordingly. This implies that said label is not as relevant in the target domain of this thesis. Hence, it was formally merged into the new `no_or_other` class again. Similarly, *collective* and *money*-related events were also merged into the same class due their low frequency in the raw dataset (see Table 8.2). The class `other_social` was a little more frequent but relates by definition to people only loosely connected to the biographical subject. As it was also difficult to distinguish such sentences from generic descriptions or non-events during the manual annotation process, it was the fourth and final original class that was merged into the residual class `no_or_other`.

Life events related to *education* were quite infrequent and proved to be very similar to the *work* domain. This is due to the fact that many biographical subjects in the corpus to be annotated worked in the field of education or were artisans, and that both fields are characterized by many events which are both educational and work-related. This is why the two original domains were merged to the new class `education_or_work` that was defined in the final annotation guidelines in the following way (see also Table A.4 in the appendix):

(8.3) *[An] event that either prepares the career of the biographical subject or is related to her or his occupation. This includes e.g. training, further education, studies, degrees, new professional activities, job loss, changes in the context of the own career.*

All these changes reduced the amount of labels from 15 "raw" ones to ten in the final manually annotated Moravian dataset. Since all alterations involved merging previous labels, the relabeling was conducted automatically.

Label	**Dataset**			
	Raw	**Train**	**Test**	**Total**
`birth*`	31	21	10	31
`death*`	78	67	11	78
`education_or_work*`	–	89	17	106
`education`	17	(15)	(2)	–
`work`	89	(74)	(15)	–
`family`	78	57	21	78
`health`	175	148	27	175
`housing`	191	149	42	191
`leisure`	39	33	6	39
`religious`	391	319	72	391
`romantic`	61	46	15	61
`no_or_other`	**988**	**839**	**221**	**1060**
`collective`	16	(13)	(3)	–
`legal`	–	–	–	–
`money`	5	(4)	(1)	–
`other_social`	51	(36)	(15)	–
Total	2210	1768	442	2210

Table 8.2: Statistics of Moravian life event domain classification datasets (labels marked with * added to the ones proposed by Haehner et al. (2024), highest support in bold, support of labels subsumed under another one in brackets)

Table 8.2 gives an overview over the label changes as well as the per-class support. It is worth mentioning that the residual class accounts for almost half of the sentences whereas *religious* events are the most attributed actual domain accounting for almost 18 %. The latter supports the lifting to a main life event domain status when analyzing Moravian memoirs.

Finally, it should be mentioned that the manually annotated sentences were once again split in randomized order into train (= 80 %) and test (= 20%) datasets with the Python module *datasets*.

8.2 Automatic annotation

In order to compare automatic life event domain classifiers to the ones applied to the other two classification tasks of this thesis, the systems listed in Section 7.2 were selected. More specifically, this means that the evaluation in the following section will again concern (i) SVMs based on different distributional features (see Section 6.2.2), (ii) the two transformer models *bert_multilingual* and

gbert (see Section 6.2.3), and (iii) the four additional zero-shot models *Gemma3*, *Nemotron_Ultra*, *bart-nli*, and *mDeBERTa-nli* (see Section 6.2.4).

Both the SVM fitting and transformer fine-tuning were based on the manually compiled train dataset presented in the previous section.

The basic prompt template shown in (6.2) was adapted for the zero-shot LLM approach by replacing the abstract task name `[C]` with *life event domain* on the one hand and constraining the answer space `[Y]` to the ten labels defined above. The resulting prompt was:

(8.4) *Please perform a life event domain classification task. Given the sentence, assign a life event domain label from ["birth", "death", "education_or_work", "family", "health", "housing", "leisure", "no_or_other", "religious", "romantic"]. Return label only without any other text. Sentence: [X] ... Label:*

8.3 Evaluation

The selection of systems capable of automatically annotating sentences from Moravian memoirs with life event domains are compared with regard to their similarities and differences to the aforementioned manual annotations in the following subsections: To this end, performance scores are discussed in Section 8.3.1. Afterwards, Section 8.3.2 offers a first analysis of selected classification errors of the best performing system, gbert fine-tuned on the task and domain at hand. This analysis is complemented by attempts to explain the general classification behavior of said system with Transformer-Explainability in Section 8.3.3.

8.3.1 Performance comparisons

A comparison of the systems selected for an automatic life event domain classification did not require any preprocessing steps because they all share the same input and output format. Like in the case of the two previously discussed tasks, the evaluation is based on weighted precision, recall and F1 score as calculated by the Python module *sklearn*. The per-system results are summarized in Table 8.3 and compared to a random baseline as well as one which only knows the majority class `no_or_other`. In the case of SVMs, the table lists only the best result per analyzer while all other are depicted in Figure 8.1. As LLM output tends to be non-deterministic, the mean and standard deviation of five different runs on five different days are included in the table.

System	Type	Precision	Recall	F1 score
Random	baseline	.25	.08	.10
Majority	baseline	.25	.50	.33
SVM_char2-3_count*	SVM*	.63	.63	.62
SVM_word1_TF-IDF*	SVM*	.67	.67	.63
bert_multilingual*	BERT*	.71	**.71**	.70
gbert*	BERT*	**.72**	**.71**	**.71**
bert_multilingual_110M	zero-shot (BERT)	.04	.12	.05
gbert_110M	zero-shot (BERT)	.28	.09	.11
bart-nli_406M	zero-shot (NLI)	.49	.19	.18
mDeBERTa-nli_279M	zero-shot (NLI)	.63	.18	.14
Gemma3_27B	zero-shot (LLM)	$.63_{\pm.01}$	$.47_{\pm.00}$	$.47_{\pm.01}$
Nemotron_Ultra_253B	zero-shot (LLM)	$.64_{\pm.02}$	$.58_{\pm.01}$	$.58_{\pm.01}$

Table 8.3: Event domain classification results (best value per score in bold, systems fine-tuned/trained on Moravian data marked with *, only the best results per analyzer are listed in the case of SVMs)

Table 8.3 shows that the best SVM settings and the two fine-tuned transformer models are quite balanced between precision and recall whereas all zero-shot models but bert_multilingual seem to have a recall issue. The latter, however, underperformed in the other two scores compared to the random baseline. Hence, the zero-shot approach is the least performant one with scores in the broad range of .04–.64 (mean .33). SVMs performed better on average with scores between .36 and .68 (mean .58) depending on the parameter setting. Finally, the two fine-tuned transformer models outperformed all other approaches with scores ranging between .70 and .72 (mean .71).

Performance of SVMs

The effect of different analyzers, n-gram ranges, and vectorizer weightings on the performance of linear SVMs trained on Moravian data is illustrated in Figure 8.1. The most impactful parameter appears to be the n-gram range with the orientation of the impact depending on the analyzer. When analyzing words, there were only minor F1 score differences between unigrams, a combination of uni- and bigrams, and a further addition of trigrams despite the different amounts of resulting vector dimensions (see Table 6.4). Yet, the removal of uni- and/or bigrams led to major performance drops. When analyzing characters instead of words, trigrams led to the best results in isolation even though the addition of bigrams led to a minimal performance gain.

The impact of the weighting scheme, however, is neglectable in most settings but led to notable performance drops when combined with single characters as well as character n-grams in the range 1–2 and 1–3.

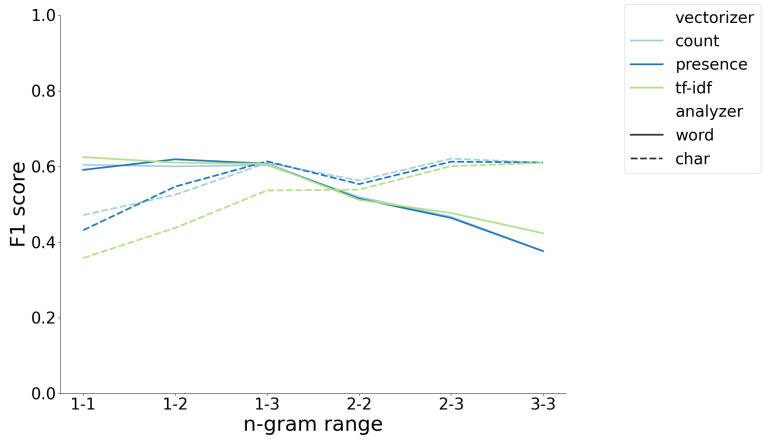

Figure 8.1: Influence of different parameter settings on SVM perfomances

Performance of fine-tuned transformer models

Both fine-tuned models were the by far most performant systems on this task although the respective base models performed close to or worse than random. More specifically, the net performance gains ranged between 44 % points in the case of the precision of gbert and even 67 % points in the case of the precision of bert_multilingual. These values underline the effectiveness of this approach even in scarce data settings. In addition, it is worth mentioning that the performance difference between the use of German or a multilingual base model was minor in the case of this specific task.

Performance of zero-shot approaches

Despite the fact that all six evaluated zero-shot classifiers had a recall issue, the precision of mDeBERTa fine-tuned for multilingual NLI as well as the two LLMs were similar to the more promising SVM settings. Moreover, there was a parameter size effect observable when looking at the weighted recall scores. Yet, even the scores of the most performant LLM Nemotron_Ultra underperformed by 7–12 % points compared to the considerably smaller fine-tuned transformer models which did not require any GPU access for inference.

8.3.2 Error analyses

As shown in the case of the two previously discussed classification tasks, a look at confusion matrices may already hint at behavioral differences between automatic annotation approaches. This is why the most promising systems per type are again evaluated in this way in the following. More specifically, Figure 8.2 depicts the confusion matrices of the best performing SVM (= one using TF-IDF weighted word unigrams as features), fine-tuned transformer model (= gbert), and best LLM run (= the third of Nemotron_Ultra).

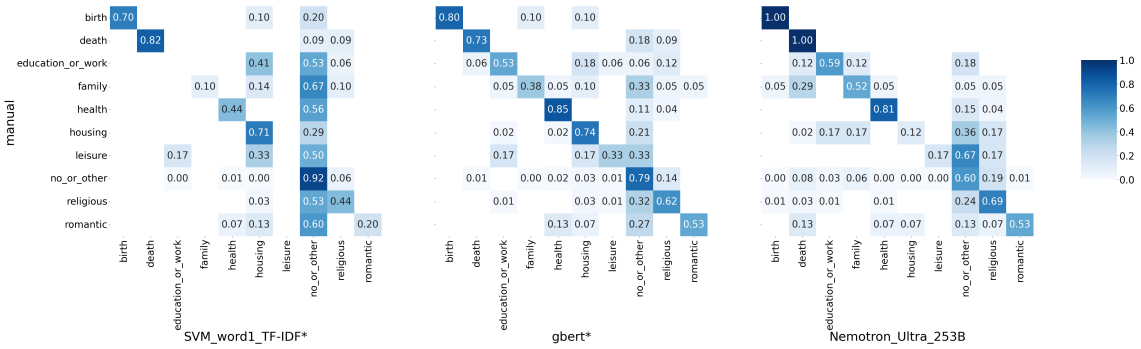

Figure 8.2: Confusion matrices of best performing event-domain classifiers per type (systems fine-tuned/trained on Moravian data marked with *)

All three classifiers were exceptionally good at identifying `birth` and `death` events implying that these classes is easier to annotate than the other major life event domains. Interestingly, both the linear SVM and even more so gbert were able to learn how to distinguish the class `birth` from the other nine classes from only 21 training samples (see Table 8.2). Nemotron_Ultra even managed to classify all `birth` and `death` instances found in the test dataset correctly in its best run but also associated a notable amount of *family* and *romantic* events with the subject's death.

All three systems had major issues with `leisure` events which may partly be due to the fact that this class was only represented by six cases in the test dataset. Hence, it is also not as relevant as some of the other life event domains in the sources under investigation. The opposite is true for the relationship between the two classes `education_or_work` and `housing`. The SVM confused the former class with the latter 41 % of the time, and the fine-tuned gbert in 18 % of the cases whereas Nemotron_Ultra showed the opposite confusion in 17 %. These confusions are not only notable but also understandable because occupational changes were commonly accompanied by a change of place in the case of 18th century Moravians (Becker-Cantarino, 2016, 176).

Finally, it should be added that all three models were seemingly biased towards the class `no_or_other`. This may be due to the fact that it accounts for almost half of the training samples in the case of the two supervised learning approaches. This confusion can, however, also be seen as some kind of recall issue as said class has a residual status.

Non-deterministic behavior of LLM classifiers and hallucination effects

This time, both LLMs yielded similar (low) standard deviations of .02 and lower (see Table 8.3) which is a result of their slight non-deterministic nature. More specifically, Gemma3 labeled 56 sentences, which is 13 % of the test dataset, differently between the five runs, and the bigger Nemotron_Ultra 51 (= 12 %). These two numbers are quite similar which is why a parameter count effect is not really observable when considering their general deterministic properties.

As observed in the other two tasks, Gemma3 added Unix line breaks in a non-deterministic fashion when labeling many instances. These were removed automatically before the performance evaluation because they were not deemed as problematic as the assignment of a completely different label. Yet, Gemma3 showed a "true" hallucination this time, too, as it introduced the new label `age` in all runs but the second in the case of Example (8.5). In the second run, it instead attributed the label `health`, which is also wrong, since the true label is `death`.

(8.5) *ihr alter hat sie auf 34 Jahr 11. Monath gebracht.*
('her Age was 34 years 11. month [sic!]')

Since similar phrases are quite common in the context of death descriptions in Moravian memoirs, it is not surprising that another one, namely Example (8.6), is part of the test dataset. Interestingly, Gemma3 labeled this almost identical sentence correctly in all five runs.

(8.6) *Sein Alter hat er gebracht auf 77 Jahr, 9 Monate u. 1 Tag.*
('his age was 77 years, 9 months a. 1 day.')

Close reading of misclassifications of the best model

Like in both previously discussed tasks, the smaller transformer model gbert outperformed all other systems when fine-tuned with the manually annotated Moravian dataset. Yet, it still produced quite a few misclassifications which will be the focus of the remainder of this chapter. A first systematic approach is to zoom in on the label confusions which are summarized in Table 8.4 by ignoring the direction of the confusion.

Confusion	Occurences
`religious` ↔ `no_or_other`	54
`housing` ↔ `no_or_other`	15
`family` ↔ `no_or_other`	8
`health` ↔ `no_or_other`	7
`death` ↔ `no_or_other`	5
`education_or_work` ↔ `housing`	4
`leisure` ↔ `no_or_other`	4
`romantic` ↔ `no_or_other`	4
`education_or_work` ↔ `religious`	3
`education_or_work` ↔ `leisure`	2
`family` ↔ `housing`	2
`health` ↔ `romantic`	2
`housing` ↔ `religious`	2

Table 8.4: Confusions of gbert fine-tuned for Moravian life event domains with at least two occurrences in the test dataset

The table illustrates that 54 confusions (= 43 %) of the fine-tuned gbert were between the two classes with the highest support in the training dataset – namely `religious` and `no_or_other`. While the confusion from the former to the latter can be seen as a recall problem, a close reading analysis of respective sentences implies that the direction of the confusion may not matter as much. Instead, quite a lot of the confusions may have actually arisen from the fact that the dataset is religious by nature. Hence, there are sentences like Example (8.7) where some kind of gratitude is expressed using *religious vocabulary*. Yet no *religious event* is described that fits the definition listed in Section 3.3.1. As the definition, however, allows for the tagging of "milestones in their faith or belief system", the difference between a taggable event and a more generic grateful or wishful phrase is somewhat granular. Therefore, it is understandable that the system struggled with this specific class on the base of textual features alone more than with others.

(8.7) *indes half mir der Heiland auch hierbei, daß ich mit ihnen durchkam.*
('meanwhile, the Savior also helped me with this, so that I got through it with them.')

The second most common confusion of the fine-tuned gbert arose between the event domains `housing` and again `no_or_other`. Here, it is noteworthy that three out of the six sentences falsely labeled as concerning *housing* changes indeed deal with such a change. However, it was in fact another person who moved to another place. In this sense, the classifier failed to account for

the fact that only events related to biographical subject should be classified with said label. This difference is likely difficult to resolve based on textual features alone without additional context. Hence, it is actually quite surprising that the model seemingly managed to factor the distinction in when classifying most other sentences. This is even more true when considering the existence of third person memoirs where a first person pronoun cannot be used as a distinguishing feature.

The three remaining errors of this type as well as nine of the reverse one are harder to motivate. The sole exception is Example (6.6) which was also an error in the sentiment classification task. Here, it is interesting to note that the tokenization issue arising from a non-standard spelling could again be the source of the confusion. Whether it really affected the algorithmic decision will be explored further in the following section.

8.3.3 Explainability analyses

When using Transformer-Explainability to assess the two exemplary misclassifications, it becomes apparent that the model indeed followed an understandable reasoning based on the subword tokens deemed relevant. As illustrated in Figure 8.3, the fine-tuned gbert model focused on the phrase *in dieser [...] lebte* 'lived in this' in the case of Example (6.6) which would be a housing description if it was not used in a figurative way. A possible explanation could be that the noun *Hof[f]nung* 'ho[p]e' is spelled in a way that makes it more similar to another housing-related term namely *Hof* 'court, yard, or farm'. This is, however, not supported by the attributed relevance scores which rather imply that this noun was deemed less important than the surrounding phrase.

text	label	birth	death	education_or_work	family	health	housing	leisure	no_or_other	religious	romantic
In dieser Hofnung lebte ich über ein halbes Jahr.	housing	1.64%	0.82%	23.44%	3.75%	1.41%	38.63%	11.62%	15.13%	2.14%	1.42%
indes half mir der Heiland auch hierbei, daß ich mit ihnen durchkam.	religious	0.11%	0.20%	0.44%	0.13%	0.83%	0.39%	0.36%	25.70%	71.39%	0.43%

Figure 8.3: Transformer-Explainability based feature relevance table for selected misclassifications of gbert fine-tuned for Moravian life event domains (see Examples (6.6) and (7.11), lighter colors correspond to less decisive tokens)

In the case of the second misclassification, it is difficult to motivate the focus on the pronoun *mir* 'me' but the second focus *Heiland* 'Savior' fits a religious motive. Interestingly, said word is also split into two subword tokens but this time in a less problematic way. The reason is that the first part *Heil* stays in the religious domain not only in the sense 'salvation' but also as a part of other words, such as *Heil[ig]* 'Holy'. It is worth underlining that gbert saw the true label `no_or_other` as the third most probable one in the former case, and the second most in the latter. Thus, it actually was on the right track but led up the garden path – so to say – by word forms humans would also rather associate with another life event domain.

Scaling local explainability scores up to the corpus level

While the analysis of two misclassifications already hinted at possible biases, an upscaling to the whole test dataset led to an even deeper understanding of the probable decision-making. To this end, Figure 8.4 depicts the overall distribution of relevance scores per subword token as computed with Transformer-Explainability. The distributions again illustrate that highly relevant subword tokens are outliers as in the case of the previously discussed classification tasks. As mentioned earlier, this also fits the expectation that most tokens are likely indiscriminatory in nature.

It is, however, rather surprising that gbert appears to have identified more discriminatory tokens for the residual class `no_or_other` than for others. This can be inferred from a higher mean relevance score on the one hand, and higher upper whiskers on the other. The same applies to the classes `health` and `education_or_work` although to a lesser degree. By contrast, the classes `birth` and `religious` led to the lowest means and upper whiskers. Whereas the former case is harder to motivate, the latter one may support the aforementioned problem of identifying textual features that distinguish religious events from other religious vocabulary.

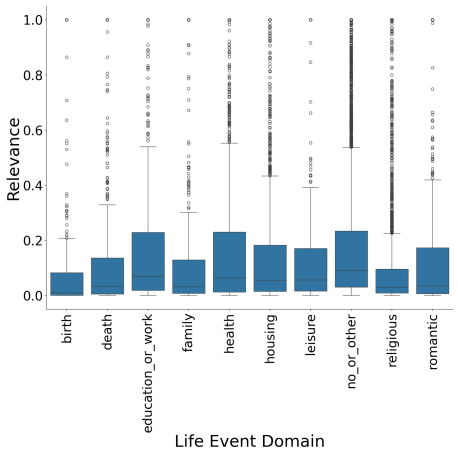

Figure 8.4: Distribution of relevance scores per class for gbert fine-tuned for Moravian life event domains applied to the Moravian test dataset

In order to investigate potential discriminatory tokens further, the most relevant subword tokens (i.e. the positive relevance outliers in the boxplots) were assessed and grouped by lexical fields (see Table 8.5). In the case of the **birth** events, there were only three highly relevant subword tokens with one of them (*geb*) being the abbreviation of another one (*geboren* 'born'). While both subword tokens fit the domain, the other one might seem a bit misplaced as *Schwester* 'sister' is not only a kinship term but also used in a more general sense of 'female Moravian' in these sources. This latter sense is actually the more probable one, here, because sentences dealing with the birth typically include this word followed by the name of the person the memoir is about. Still, it is striking that the male equivalent *Bruder* 'brother' is deemed less relevant for the prediction of this class. Furthermore, it is surprising that gbert has so little problems with predicting the class `birth` when considering its confusion matrix (see Figure 8.2) given that it has only seen 21 training samples (see Table 8.2). This is likely the reason why it could not learn more highly relevant subword tokens. Another one could very well be the fact that birth descriptions are rather formulaic in Moravian memoirs which is a feature the model learned to exploit.

When disregarding the different amount of samples in the training dataset, most of these findings also apply to descriptions of the subject's **death** which is not so surprising as this event domain is the natural opposite. The most relevant subword tokens hint at typical formulaic expressions around the lexemes *begraben* 'buried' on the one hand and *Alter* 'age' on the other. The latter lexical field also includes a number which fits the custom of specifying the current age and/or year right after the first mention of the death in these sources. So, while six of the most highly relevant subword tokens relate to the target event domain, the remaining two *##la* and *ents* are in fact too ambiguous without their context.

Considering the merged domain **education/work**, most of the subword tokens gbert appears to regard as highly relevant are definitely relevant for this class. They are (i) synonyms of the word 'work' like *Arbeit*, *Profession*, or *Dienst*, (ii) typical profession compounds ending with *##meister* 'master', or (iii) places like *[Chor]häuser* '[choir] houses' where Moravians typically worked. Yet, it is remarkable that the focus seems to be more on the domain *work* than *education*. This may be explained by the distribution of these two domains which were distinct during the original manual annotation but merged due to lacking support of the latter domain. In fact, the rather low support resulted in only two instances from the latter (sub)domain in the test dataset (see Table 8.2) so that it is difficult to assess whether this aspect of the merged class was learned, as well.

A domain merging problem is not present in the case of the **family** events but there are again only a few highly relevant subword tokens. Yet, basically all listed words fit the domain very well because they are kinship terms. The word *Ehe* 'marriage' is the only outlier and is a little more difficult to motivate because it may seem more apt for the domain `romantic` where it is in fact also listed. In combination with the earlier observation that both classes were never confused with each other, one can infer that the model may have learned to differentiate between marriages of the

subject on the one hand and of other family members on the other. This would be a major learning success as such a distinction may seem rather difficult based on textual features alone.

In principal, it could also be difficult to assess whether events related to **health** concern the biographical subject or someone else. However, the Moravian memoirs under investigation rarely include health information about other people. In this sense, it can be seen as a learning success that highly relevant subword tokens from the lexical field 'health' are not found in other per-class lists. The sole exception is *Leiden* 'suffering' listed under the domain `religious` which can be explained with the many references to Christ's suffering. Regarding other tokens deemed relevant for the class `health`, the occurrence of body parts like *Kopf* 'head' and *Bein(e)* 'leg(s)' are noteworthy. Their high relevance supports the assumption from Section 7.3.3 that gbert may have learned to associate body parts with health issues and in extension with an old age when fine-tuned for a life stage classification. While this letter step may be a little more far-fetched, the first one seems plausible. By contrast, it is not so clear how to rate the actual relevance of most of the tokens categorized as being ambiguous – mostly because they are actual subword tokens and no words. Nevertheless, a closer inspection revealed that some of them co-occur. Two examples of that kind are *schw+ächlich* 'weak' and *schw+ollen* 'swollen' both of which are relatable to health.

Only a few subword tokens were deemed relevant for **housing** events but most if not all of subword tokens found by Transformer-Explainability fit the respective class description. This is especially true for the movement describing verbs *zog(en)* 'moved' and *kam* 'came', and maybe also for the word *Bestimmung* 'determination'.

Interestingly, the closely related lexical field of locations represented by *wo* 'where' is one of only three discriminatory subword tokens for **leisure** events. While the low amount may be caused by the comparably small amount of training and test data instances, it is remarkable that the other are *Besuch* 'a visit' and *besuchen* 'to visit'. As a matter of fact, leisure activities may encompass more aspects in general but a close reading implied that all test instances actually dealt with visits. Hence, an explainability approach based on these instances can just not asess whether the model may or may not have learned to identify other leisure activities.

Residual classes are, in general, more likely inferred ex negativo as discussed in Section 6.3.3. Yet, Transformer-Explainability found more highly relevant subword tokens for **no or other** events than for any other event class – albeit that most of them are ambiguous in isolation. Thus, it appears to be the case that gbert was able to identify some typical aspects of this specific residual class. Most notably, this applies to the identified occupation names and military vocabulary since they are relatable to the respective subdomains `other_social` and `collective`. Furthermore, especially locations are commonly found in generic descriptions so that their occurrence in the list of relevant tokens is comprehensible. However, tokens from the lexical fields of locations, movements and even more so kinship terms may also explain the aforementioned "recall" problem for the classes `housing` and `family` (see Figure 8.2).

Interestingly, the latter applies to a lesser degree to the most common "recall" issue where instances of the domain `religion` are falsely labeled as such since only the two profession names *Pastor* and *Pfarrer* 'pastor' are relatable to said class. These are, however, not deemed as relevant for **religious** events as gbert has seemingly learned to factor in other typical words. The most relevant tokens for said class include *Gebet* 'prayer' and *aufgenommen* 'accepted [to the community]'. Similarly, it is possible that *Bitte* 'request/rogation' and *besuchen* 'visit/attend' were used in a secondary (more religious) sense in respective sentences. If so, this would suggest that gbert was able to grasp such differences. In addition, there is the word *Hände* 'hands' which could refer to 'Christ's suffering' (Faull, 1992, 37). Finally, the two pronouns *Ich* 'I' and *Ihn* 'Him' may be part of expressions of gratitude towards the Savior as suggested by both the case and casing of these word forms.

Last not least, the list of most relevant tokens for **romantic** events is again a short one but comprises only subword tokens which are the word *Ehe* 'marriage' or part of words like *Trau[ung]* 'wedding' and *verhei+rath[en]* 'to wed' which fit thematically.

In summary, this means that the fine-tuned gbert model predominantly considers token features predictive a human being would also deem relevant which is a remarkable learning success.

8.4 Summary

While it may seem anachronistic to apply a modern-day categorization of major life event domains to historical sources, manual annotations underlined that most categories proposed by Haehner et al. (2024) were in fact generic enough for this kind of domain adaptation. Still, a few classes of their hierarchy needed to be refined in order to account for peculiarities of the biographical data under investigation here (see Section 8.1). First of all, the two new classes `birth` and `death` were introduced. Secondly, it was necessary to merge the original classes `education` and `work` because many biographical subjects worked in the educational system. Last not least, the four classes `collective`, `legal`, `money`, and `other_social` played such a minor role in this corpus that they were merged into the residual class `no_or_other`.

Interestingly, this latter class was annotated the most even when disregarding the additional instances merged into it. `Religious` events had the second highest support which is due to the fact that the respective biographical sources are religious in nature.

Concerning automatic annotation approaches, (i) SVMs fitted to distributional features, (ii) the LLM Nemotron_Ultra, and (iii) fine-tuned transformers yielded the most promising results in this ascending order (see Section 8.3.1). The **SVM** approach had a striking issue with discriminating the more important domains from the class `no_or_other`. The **LLM** Nemotron_Ultra performed better with regard to some classes but confused more with each other (see Section 8.3.2). It also showed notable non-deterministic behavior by classifying 12 % of the test instances differently between different runs. This also applies to the smaller LLM Gemma3 which was a little more non-deterministic and halucinated the label `age` in one case. By contrast, the **fine-tuned transformer** gbert and its multilingual counterpart were both more balanced with regard to their per-class and overall performance. Thus, said supervised learning approach resulted once again in the most promising system(s) for the task and domain at hand.

Although gbert still misclassified a notable amount of test sentences after having been fine-tuned, many of these errors can be motivated by either a close reading analysis (see Section 8.3.2) or a look at relevant subword tokens identified with Transformer-Explainability (see Section 8.3.3). The most common issue was that many phrases include religious vocabulary although they do not necessarily describe events from this domain. Similarly, the model sometimes failed to identify whether the biographical subject or another person moved to a new place. Nevertheless, Table 8.5 underlined that gbert has learned to identify many typical words for basically every life event domain during fine-tuning no matter the amount of training samples. Especially in the case of the rather rare class `birth`, it probably exploited the typical formulaic phrasing involved. In summary, this means that said transformer model has learned to not only distinguish between the different proposed life event domains in general but also to account for peculiarities of the target data.

Label	Subword Tokens	Lexical Field
birth	Schwester	*ambiguous*
	geboren, geb	birth
death	Alters, Geburts, Alter, alter, 71	(old) age
	begraben	death
	##la, ents	*ambiguous*
education_or_work	##meister, Arbeits, Profession, Dienst	occupation
	Hier, Chor, ##häuser	location
	des	pronoun
family	Mutter, Schwester, Söhne, Söhnen, Tochter, Enkel, Vater, Töchter	kinship term
	Ehe	relationship
health	Gesundheit, krank, besser, Kranken, schwach, Schmerzen	health
	##wollen, schw, fr, ##fie, ##werden, ##heit, ##k, ##ächlich, ##besch, ##uls, Krä, ##klich, ##än, ##chen, bef, nehmen, Schw, Seiten, ##keit	*ambiguous*
	Kopf, Beine, Bein	body part
housing	169	date/number
	Bestimmung	*ambiguous*
	dahin, zog, zogen, kam	movement
	dieser	pronoun
leisure	wo	location
	besuchen, Besuch	visit
no_or_other	##41, Date, 179	date/number
	Pastor, Arbeit, Arbeiter, Pfarrer	occupation
	Worte, Versammlung, Sachen, Person, Nachrichten, Daran, Nach, Magd, Luther, ##aden, Zit, angebaut, art, begeg, bitte, brennen, einzurichten, empfangen, erworben, folgendermaßen, geschlagen, gewünscht, kennen, singen, solche, verkauf, warme, ##q, Brief, ##lager, ##sagen, ##fus, ##fin, ##fest, ##ühle, ##ferenzen, ##erbe, ##einander, ##egen, ##aus, ##atur, ##liebe, Rauch, Leuten, ##tw, Jungen, ##ommen, Auflösung, ##inn, ##ren, ##bst, Unglück, fragte, Nachbarn, höchst, ##indern, Singen, schreibt, ##iod, ##uh, Von, Spruch, ##chte, Kleider, ##mün, jens	*ambiguous*
	Wache, Krieg, Führung	military
	Schiffes, fahren, Bewegung, Lauf, lief	movement
	Hütte, Häuser, ##häus, Berg, heim, Anwesen	location
	Tochter, Mutter, Vater, Bruder	kinship
religious	Gebet, Sel, aufgenommen, Geist, Glauben	religion
	Bitte, besuchen, Ves, ##tu	*ambiguous*
	Hände	body part
	Leiden	health
	Ich, Ihn	pronoun
romantic	Ehe, Trau, verhei, ##rath	relationship

Table 8.5: Subword tokens with a minimum relevance of .75 for gbert fine-tuned for Moravian life event domains

Chapter 9

Biographical content analyses

Having described the three classification tasks of this thesis in the previous chapters, this chapter deals with three possible application scenarios from biographical research which are all content analyses. Section 9.1 outlines the methodological approach they follow before Section 9.2 shows patterns found in 18th century Moravian memoirs by first considering the classification tasks separately and subsequently relating them to each other. Section 9.3 builds up on this with an investigation of one demographic variable – the gender of the (auto)biographical subject. Finally, Section 9.4 highlights differences to patterns found in another biographical subgenre – namely entries in the 19th century biographical dictionary ADB.

9.1 Experimental setup

As noted in Section 2.2, distant reading approaches can be seen as experiments in a social science sense meaning that hypotheses, methods, and samples are defined a priori. To this end, Section 9.1.1 summarizes the most important hypotheses derivable from related work. Section 9.1.2 illustrates the methodological approach. It hence shows how models from the three classification experiments outlined in the previous chapters can be reused for downstream tasks that aim at a quantitative verification or falsification of results from earlier studies which were mostly qualitative in nature. Finally, Section 9.1.3 explains the genesis of the different samples used in the exemplary content analyses outlined in the subsequent sections of this chapter.

9.1.1 Hypotheses

Hypotheses that can be tested within content analyses based on sentiment, life stage, and life event domain classifiers arise from typical features of biographical sources (see Section 2.1.1), major life events (see Section 3.3.1), Moravian memoirs (see Section 4.1.2), and ADB entries (see Section 4.2.1). They can be grouped in the following way:

Hypotheses regarding life events in general

- There should be more positively connoted life events in young adulthood than later in life (Haehner et al., 2024, 11–12).
- There should be more health events in old age than earlier in life (Haehner et al., 2024, 10).

Hypotheses regarding Moravian memoirs

- Memoirs should follow the life course chronologically in general (McGuire, 2021, 132) albeit that the retrospective sense-making involved may lead to non-linearities (see Section 3.2.1).
- There could be individual differences in the depiction of life courses (Faull, 2021, 217–218).
- Memoirs should show a high degree of emotionality (Van Gent, 2017, 245).

- Conversion narratives should show a gradually rising polarity (Elkins, 2022, 105–107).
- McGuire (2021, 133) hypothesized that the polarity found in the memoirs should ...
  - "begin as neutral"
  - rise "to a general positive peak in early childhood"
  - "fall through adolescence and young adulthood"
  - "rise slightly as people describe the conversion experience"
  - "decrease [...] but remain positive overall" during midlife
  - "rise sharply in a final peak"
- The depiction of final illnesses and death can sometimes be longer than the rest of the text (Schmid 2004, 51; Böß 2016, 238) and should be positive (Faull and McGuire, 2022, 143).
- Since romantic events like weddings were the result of lot drawing in some 18th century communities (Becker-Cantarino, 2016, 185), they could be more neutrally connoted.

Hypotheses regarding gender differences in Moravian memoirs

- There should be no apparent gender differences (Mettele, 2007, 33–34).
- Faull and McGuire (2022, 144) showed that "women express[ed] more positive emotions than men" in 18th century English memories which should also be true for German ones.

Hypotheses regarding genre differences between memoirs and dictionary entries

- Memoirs should be more anecdotal and less total (see Section 2.1.1).
- Memoirs should have a more "incremental, episodic structure" (Buss, 2001, 595).
- Autobiographical records like memoirs should focus more on youth and biographical ones like dictionary entries more on later life stages (Cockshut, 2001, 78).
- Dictionary entries should be more neutral in tone in general albeit that there may be signs of discipleship and/or hostility (Cockshut, 2001, 79).
- ADB entries may start with demographic facts (Reinert et al., 2015, 16) and a genealogical abstract or a summary of the most important achievements (Reinert and Scholz, 2017, 5).
- ADB entries are not required to follow hypotheses regarding Moravian memoirs.

9.1.2 Analytical methods

Many hypotheses relate to trends over the text course (= narrative time) and/or the actual life course (= narrated time). This is why the methodological approach of the content analyses to follow is centered around said aspects:

Polarity trajectories

Chapter 6 showed that some Moravian memoirs have been manually annotated with sentiment labels and that these annotations can be used to train classifiers able to automatically annotate more biographies – or sentences to be more specific. The **sentence tokenization** was again implemented with stanza and followed by an automatic merging of "sentences" with date splits (see Chapter 5). In order to gain trajectories, the categorial sentiment labels were **mapped** to their typical valence score counterparts – i.e. `negative` was mapped to `-1`, `neutral` to `0`, and `positive` to `1`. Since the biographies differ in length, each one was **binned** into 10 % steps and each bin was attributed the mean valence score. After aggregating, the neutral range was expanded to cover `[-.05,.05]` as is common practice (Hutto and Gilbert, 2014, 224).

Life stage and event domain distributions

While sentiment labels can be mapped to numerical values, life stage (see Chapter 7) and event domain labels (see Chapter 8) are more categorial in nature. This is why they were analyzed slightly differently with a stronger focus on the relative distribution of sentences dedicated to each individual label. This will be first related to the text as a whole and subsequently binned in order to analyze distribution changes over narrative time.

Chronological rescaling to the narrated time

The 10 % step bins account for the narrative time whereas, for instance, the hypothesized polarity trends of Moravian memoirs relate to the actual narrated time. To resolve this mismatch, the life stage labels were subsequently used as another grouping variable besides the bins by again aggregating to mean polarity scores. In a similar vein, the life event domain distributions were also rescaled from bins to life stages.

Depiction of life event domains

Having related sentiment polarity and life event domains to life stages with the chronological rescaling, they were also related to each other as a means of assessing whether some domains are depicted more positively or negatively than others. The exact event domain labels have, to the best of my knowledge, not yet been used in other digital humanities studies. However, a depiction of events and similar topics definitely have (see Section 3.1.3) so that further contextualization possibilities arise. In addition, the relationship between life stages, event domains, and sentiment polarity can be related to modern day biographical research from the social sciences where *sentiment polarity* is typically called *valence* (see Section 3.3.1).

9.1.3 Resampling strategies

The data base of the content analyses to follow comprises various subsamples of the biographical texts described in Chapter 4. First of all, manually annotated Moravian memoirs are compared with predictions of fine-tuned gbert models in Section 9.2. The memoir corpus is split by gender (see Section 9.3) in order to account for the influence of this demographic variable which has also been investigated in related work. Finally, Section 9.4 exploits the fact that textual embeddings of gbert are quite similar between Moravian memoirs and ADB entries (see Figure 4.2) so that an application to the second corpus seems possible. This further upsampling enables an assessment of subgenre differences but is accompanied by the issue that some ADB entries were too short for the binning. Thus, only entries with more than ten sentences were considered which reduced the ADB sample to 2630 (= 74.1 %). The final corpus statistics are summarized in Table 9.1.

Dataset	Annotation	Documents	Sentences	Tokens	Types
MM	manual	36	2,210	55,285	10,129
MM	automatic	89	6,473	137,092	63,702
ADB	automatic	2,630	115,486	2,903,865	245,003

Table 9.1: Basic corpus statistics of analyzed datasets

9.2 Distributions in Moravian memoirs

This first exemplary application scenario of the sentiment, life stage, and event domain classifiers described in the previous chapters closely follows the methodology outlined in Section 9.1.2: First of all, the manual annotations of each task are considered individually and upscaled to the full Moravian memoir corpus (see Section 9.2.1). In order to investigate trajectories over the text course,

they were kept in the original order instead of being randomized like in the classification experiments. Section 9.2.2 then shows how a scale change to the narrated time affects the conclusions drawn from an analysis of the narrative time alone. In this way, it is possible to validate the hypothesis that Moravian memoirs mostly follow the life course chronologically. In addition, Section 9.2.3 investigates the sentiment polarity of life event domains and hence completes the binary task combinations. Afterwards, Section 9.2.4 summarizes the main findings of this content analysis and leads over to the subsequent analysis of gender differences.

9.2.1 Distributions over narrative time

Since no classification experiment yielded perfect performances, annotations of the best model (= gbert fine-tuned on Moravian instances) were compared to manual ones in order to validate an upscaling to the full corpus. This resulted in three samples per task which will be compared to each other in the following in order to identify notable distribution patterns over the narrative:

Sentiment polarity distributions and trajectories

English 18th century Moravian have already been analyzed with regard to their sentiment (McGuire, 2021; Faull and McGuire, 2022). This is why said task was chosen as a starting point for the analysis of German ones, here, as well. Looking at overall distributions on the text level (see Figure 9.1), it becomes apparent that (i) these sources are rather positive and (ii) the differences between manual and automatic annotations are minor. The second finding implies that the slight bias towards the class `positive` of gbert (see Figure 6.2) mostly averages out when aggregating. Still, this classification behavior could be part of the reason why the automatically annotated full sample of 89 memoirs appears to be even more positive. In addition, the high degree of positivity found in all three samples also fits the assumption of a highly emotional life depiction, as asserted by Van Gent (2017, 245).

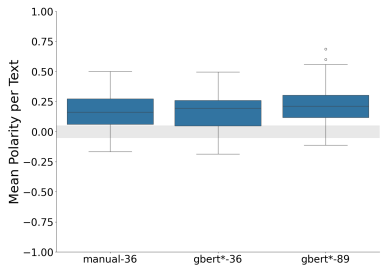

Figure 9.1: Distribution of sentiment polarity in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

When considering the text course in narrative order, the negativity becomes even less effective as illustrated in Figure 9.2. As a matter of fact, all three trajectories are quite similar in the sense that they show an overall rise in polarity over narrative time. This matches the hypothesized feature of conversion narratives (Elkins, 2022, 105–107) and can be related to the emotional arc called "Rags to riches" by Reagan et al. (2016). More specifically, the trajectories start in the neutral range and end with a global peak. Both findings fit the expected neutral demographic section at the beginning and the positively connoted "going home" moment in the end (see Table 4.1). In addition, it is notable that especially the fully upscaled version stays in the positive range over the whole narrative time which again matches the assumption of being highly emotional.

The main differences between the three versions regard the first and the second third of the narrative. The first third seems to be depicted mostly neutral or even slightly negative in the smaller sample while including a first minor peak in the case of the full corpus. In a similar vein, the subsequent second third of the narrative involves more ups and downs in the 36 memoirs than in all 89 where a second minor peak is observable. However, this part is still mostly positive in the smaller sample, too. This minor difference can be seen as a reduction in harmonicity or the number

of oscillations (see Figure 3.5). It also leads to a less pronounced low at the end of the second third of the narrative in the largest sample. Still, all three versions are in line with the hypothesis that "the polarity should decrease [...] but remain positive overall" (McGuire, 2021, 133).

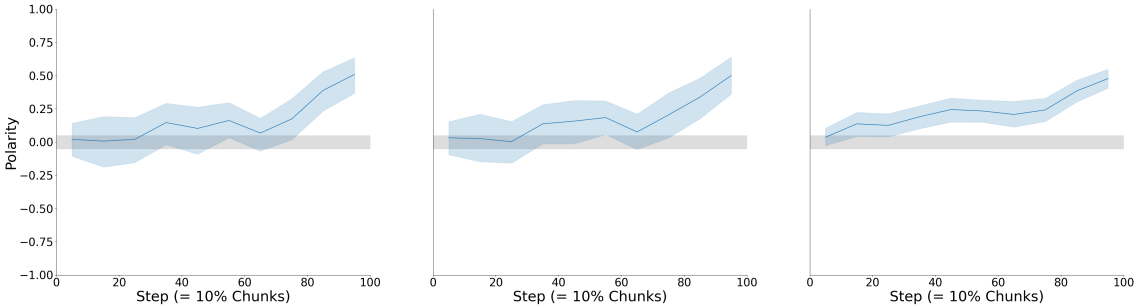

Figure 9.2: Sentiment polarity trajectory in Moravian memoirs over narrative time (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

Life stage distributions

With respect to life stages, it was already shown in Section 7.1 that the *(early) adulthood* stages predominate in the manually annotated sample. In fact, this distribution becomes even more skewed when using gbert (see Figure 9.3). This is due to its imperfect identification of sentences which describe experiences during childhood and adolescence so that this bias should be kept in mind. Nevertheless, the relation between the attributed labels stays quite consistent when upscaled. Hence, the focus on later life stages may relate to the ascribed value of viewing final illnesses "as a trial to bear" (Faull and McGuire, 2022, 143).

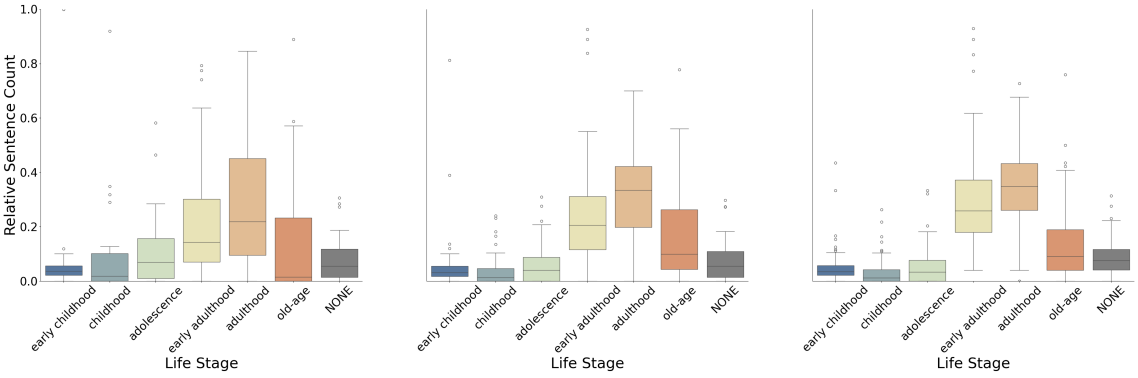

Figure 9.3: Distribution of life stages in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

The distribution of life stage labels over the narrative time is depicted in Figure 9.4 and illustrates that these sources indeed follow the life course rather chronologically. For instance, the mean amount of sentences dealing with *early childhood* experiences drops after the first 10 % of the narrative whereas descriptions of an *old age* increase over narrative time in all three scales under investigation. This also applies to the intermediary life stages so that the underestimation of childhood and adolescence proportions somewhat averages out.

Still, it should be highlighted that most bins comprise larger proportions of earlier and later life stages than expectable from a strict chronological order. This could be due to some Moravians writing more text about earlier stages and others focusing on later ones so that these foci overlap in a corpus-wise overall distribution. One reason for differences of this kind is that some people died so earlier that their memoir just cannot comprise sections about later life stages. A different and also not completely unlikely interpretation of this finding is that pro- and analepses cannot be

completely ruled out. This would be in line with biographical research stressing that such stylistic devices are used for retrospective sense-making purposes (see Section 3.2.1).

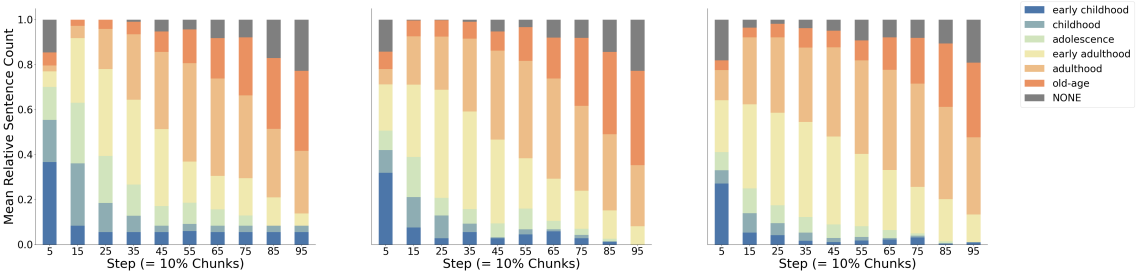

Figure 9.4: Distribution of life stages in Moravian memoirs over narrative time (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

Finally, it is an interesting finding that *non-biographical remarks* predominate in the beginning and increase over the last third of the narrative as this fits their expected framing purpose. However, there is also a certain correlation with the final life stage so that it seems as if choir helpers posthumously added more comments on recent events than on more distant ones. This is somewhat plausible given that these events were more present to these (co-)authors at the time of writing.

Life event domain distributions

Considering the overall distributions of life event domains (see Figure 9.5), it is worth mentioning that there were again only minor differences between the manually annotated memoirs on the one hand and the automatically annotated samples on the other. Similar to the distribution described in Section 8.1, most memoirs comprise notable amounts of sentences that do not describe biographical events as defined in this thesis. This appears to be even more the case in the full sample even though it is unclear if this finding is a side effect of the recall issue analyzed in Section 8.3.2.

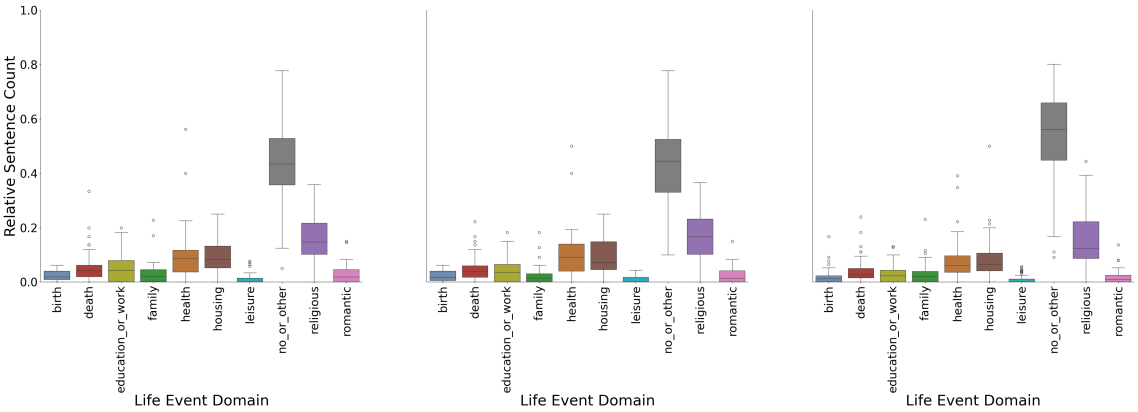

Figure 9.5: Distribution of life event domains in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

The second most common event domain is the *religious* one which is not so surprising because these sources focus on the "religious self" of the subject as Mettele (2007) put it. Interestingly, *health* and *housing* events follow which underlines that these are important topics in these sources. The former can be related to the aforementioned emphasis on final illnesses and the latter to the fact that the lives of Moravians involved many housing changes in the 18th century due to their missionary work (see Section 4.1.1). As a matter of fact, this is also in line with the high amount of travel descriptions among 18th century Moravian sources mentioned in the same section.

The potential recall issue is visible in the binned version which takes narrative time into account as well (see Figure 9.6). However, it is less problematic than one could assume as it mainly affects birth events in the first bin and death as well as romantic events in the fifth while all other cases

seem to average out. Hence, the following patterns are typical for the samples under investigation: The *birth* of the subject is almost exclusively described in the beginning and almost never taken up in later parts of the narrative. By contrast, the *death* is sometimes already mentioned in the beginning and plays an increasing role over the last third of the narrative. Its occurrence in the very beginning relates to the funeral setting. This explains why some memoirs start with the notion of the recent death whereas the final increase once again fits the importance of the "going home" moment. In addition, this second finding confirms the observation that the final section, which prototypically deals with illnesses and death, may span the largest part of the narrative (Schmid 2004, 51; Böß 2016, 238).

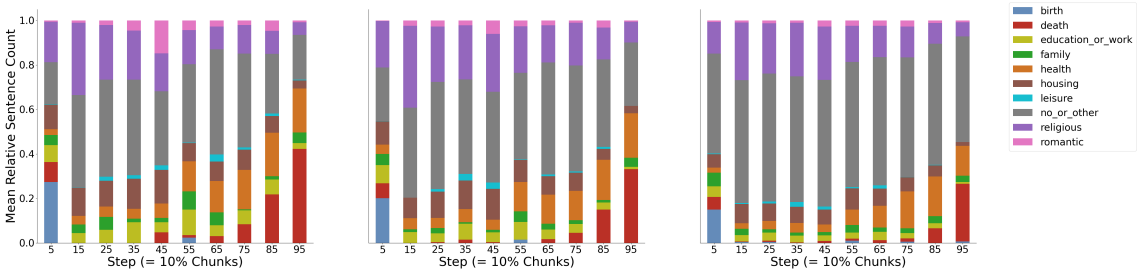

Figure 9.6: Distribution of life event domains in Moravian memoirs over narrative time (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right)

Unlike birth and death, events belonging to other domains can occur at any time in life in theory. This is why one could assume that their share of narrative time stays basically the same no matter the bin when disregarding minor variation. As a matter of fact, this applies, however, only to *family* and *housing* events.

In the case of the merged domain *education/work* it is worth highlighting that they are distributed quite evenly over all bins but the last. This could either be motivated by the probability of retirement – or at least a workload reduction. Another interpretation could once again be the focus on final illnesses Moravians counted as struggles one should overcome in order to reach salvation. The latter can also be inferred from the distribution of *health* issues and improvements since they are not only described in more detail than events related to education or work but also occupy an increasing proportion over narrative time.

The proportion of the *religious* event domain decreases which may relate to the fact that some events that fall in this category, such as baptism or confirmation, usually happen in earlier life stages. It was also not uncommon that people joined the Moravian church as young(er) adults at that time. Hence, it is less likely although still possible that other religious events which occurred later in life had an as lasting impact on the belief system of the subject that they were deemed worth including in the respective memoir.

Last not least, it should be added that *leisure* and *romantic* events tend to be described in more detail in the middle of the narrative although being actually rather rare in total. In this sense, they appear to be less important than other domains in these sources.

9.2.2 Distributions over narrated time

Given that (i) some hypotheses rather relate to the life course and not necessarily the text course, and (ii) the former is only approximated coarsely by the latter (see Figure 9.4), distributions over narrated time are considered next. Since the life course equals narrated time in the case of (total) biographies and can be discretized into life stages, labels from this second task allow for such a chronological rescaling. However, this approximation also implies that only polarity trajectories and binned life domain distributions but no life stage distributions can be assessed in this specific methodological manner.

Polarity trajectories

The polarity trajectory over narrated time was shown in Figure 9.2 and its chronologically rescaled version which combines the sentiment annotations with life stage labels is depicted in Figure 9.7. Different from the first version, this one shows more differences between the manual and automatic annotations – first and foremost with respect to the depiction of childhood and adolescence. As mentioned earlier, these life stages were commonly confused with (early) adulthood by gbert so that some of the differences may be a side effect of the imperfect classification.

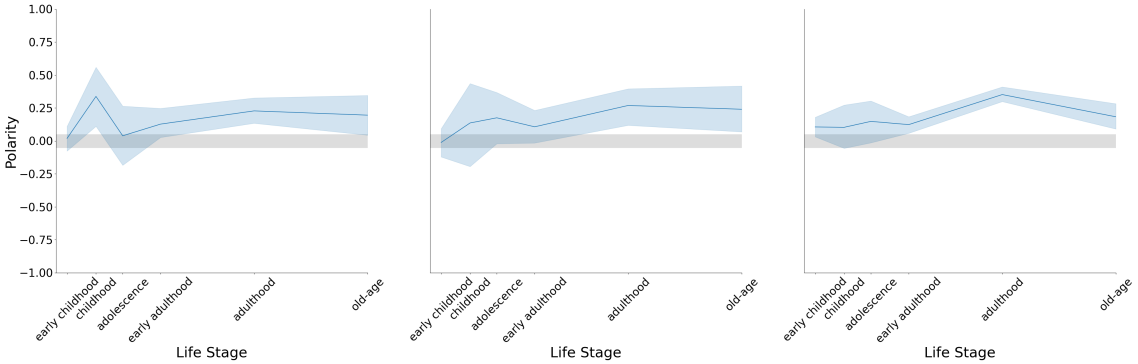

Figure 9.7: Sentiment polarity trajectory in Moravian memoirs over the narrated time (36 manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right)

One such difference is the hypothesized *major childhood peak* which is only reflected in the manual annotations. The respective automatic annotations, by contrast, are rather characterized by notable variation while the peak is lacking in the full sample. Interestingly, all three versions still include a (slight) rise – especially when looking at the overall mean – which becomes less apparent and more delayed into subsequent life stages during upscaling. Given the known biases of the classifier, this finding may not completely falsify the existence of the presumed childhood peak. This is even more true when considering that it only emerged in the case of the manual annotations after rescaling to the narrative time. Thus, another possible reason could be an overlap with the hypothesized subsequent "fall through adolescence and young adulthood" (McGuire, 2021, 133) as both these life stages cover large parts of the first three bins (see Figure 9.4). As a result, both presumed patterns could average out when looking at the polarity of the initial narrative section. They even diminished in the case of a classifier which tends to have issues distinguishing between these specific life stages – maybe because they were less distinct from each other in the 18th century. A potential further side effect of this issue is that the low during adolescence which is clearly observable in the manual annotations is diminished as well and postponed into early adulthood when scaled up.

The subsequent second peak which falls into adulthood prevails the scale change and becomes in fact even more pronounced in the case of the automatically annotated full sample. The fact that the differences between the manual and automatic annotations are minor in this case underlines that the model is better at predicting these later life stages. Hence, this is a remarkable finding – even more so because it contrasts not only the trajectory over narrative time but also the presumed midlife low (see Figure 9.8).

In a similar vein, the upscaling of the trajectory over narrated time leads to the emergence of a *final fall* because the depiction of old-age stays consistent while adulthood becomes more positive. This fall contradicts the "going home" motive and hence the overall polarity trajectory expectable from a conversion narrative which should have a clear positive outcome. When related to prototypical emotional arcs (Reagan et al., 2016), this means that the idealized one follows the "Cinderella" pattern whereas the actual narrative trajectory found here matches the "Rags to riches" arc and its narrated counterpart the "Icarus" arc. Hence, one could claim that the depiction of the narrative differs typologically from the underlying life course so that especially this difference in the end is worth investigating further. One possible reason could be the observed co-occurrence of old-age experiences and non-biographical remarks. This could imply that the addition of the latter

aimed at fitting the expectation of a more positive end. Another possible explanation is that this finding confirms the hypothesized field of tension between the negativity of illnesses and eventual death and the positively connoted "going home" motive (McGuire, 2021, 133). In fact, the former is more related to the narrated life course whereas the latter can be seen as a more narrative feature so that it is actually not so surprising that these differences become (only) apparent when changing the analytical perspective.

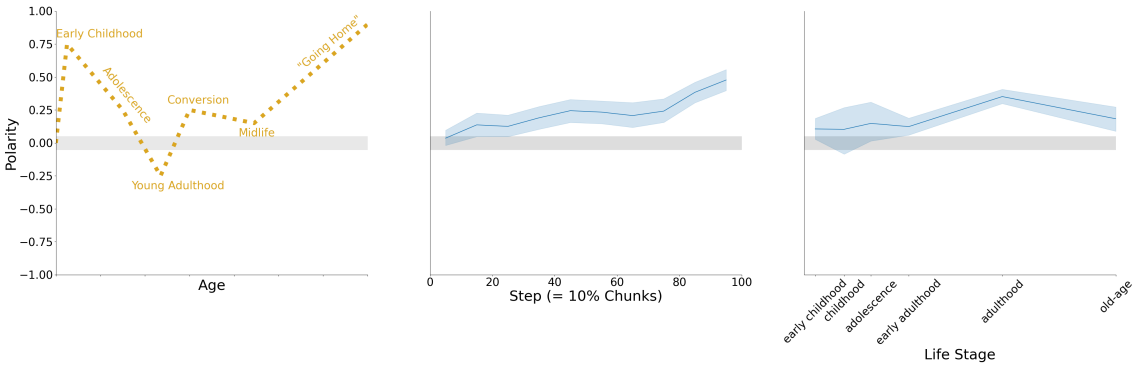

Figure 9.8: Comparison of Moravian sentiment trajectories (idealized based on McGuire (2021, 133) left, 89 memoirs predicted by fine-tuned gbert models center, and rescaled to the narrated time right)

Life event domain trajectories

When rescaling life event domains to the narrated time (see Figure 9.9), it becomes apparent that *birth* really only occurs during early childhood as expected. This implies that the slight proportion of birth descriptions at step 55 observed in the distribution over narrative time is likely a side effect of the not completely linear life description or related to rare analepsis cases. The distribution of the contrary event *death*, however, is probably better explained with life expectancy and mortality rates: For instance, the age at death distribution of the analyzed biographical sources (see Figure 4.4) fits the observable decrease over the first three life stages and subsequent increase over the later ones. Moreover, there was a high infant mortality rate in the 18th century in general which explains the notable proportions of death-related passages during (early) childhood in the manual annotations. However, it should also be noted that the fine-tuned gbert model failed to identify these (early) childhood deaths – maybe because death descriptions are textually similar no matter the life stage involved and tend to occur more often later in life. Hence, although this bias is understandable, it may be the reason for the lack in the upscaled version, too. This is why said difference between the two samples is difficult to interpret.

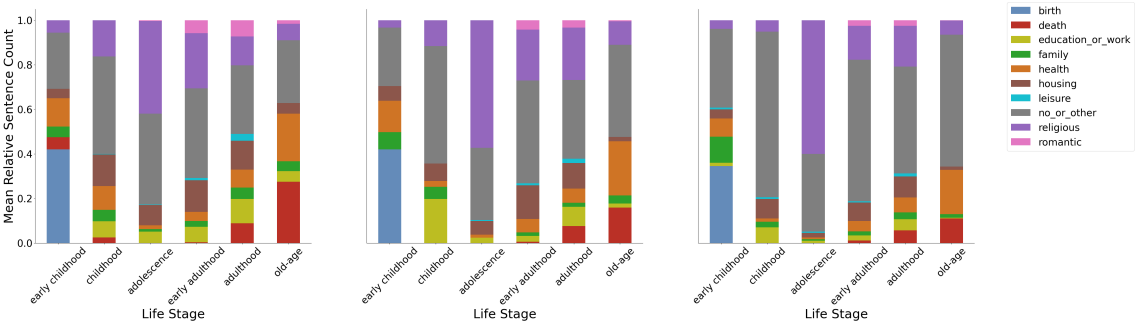

Figure 9.9: Distribution of life event domains in Moravian memoirs over the narrated time (36 manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right)

The increase of *health*-related events can be interpreted more easily as there is again a co-occurrence with the distribution of death descriptions. Since the classifier had in fact the least

issues with identifying health events, one could even see the comparably high proportion during early childhood as some kind of indication for the existence of child death descriptions in the larger sample. The correlation itself is remarkable and most likely genre-specific because health issues can be non-lethal in theory. Yet, the distributions imply that mostly only lethal illnesses play a role in Moravian memoirs. Hence, the observable increase over the narrated time is actually not so surprising since lethal health issues become more likely as people become older.

Considering the merged domain *education/work*, it is worth highlighting that related events are absent in the smaller sample and really rare in the upscaled one in early childhood sections. Similarly, they account for a smaller proportion during old age than in the case of the four intermediary life stages where they are more equally distributed. This becomes even more true when factoring in that gbert overestimated said event domain in childhood descriptions. This overall distribution fits the general assumption that people go to school as soon as they reach childhood, start working during adolescence or early adulthood, and work less when old.

In a similar vein, it is possible to explain why *housing* events are distributed the way they are since it is less likely that people move to a new place when they have young children or are of old age. The same is true for the occurrence of *leisure* activities which are not only quite rare but also mostly restricted to (early) adulthood in these sources.

Regarding subdomains of social events, it is noteworthy that *family* and *romantic* events play almost no role during descriptions of adolescence. More specifically, the former are almost equally distributed over the other life stages while the latter are restricted to (early) adulthood experiences. Moreover, especially the fully upscaled version implies that family events focus on parents and siblings as they mostly appear during descriptions of (early) childhood. The life stage restriction of the rarely mentioned romantic events is also interesting as it is understandable that these do not fall into (early) childhood but worth highlighting that there is also a lack during adolescence where first notable romantic events could occur in theory. This lack may be explained with the predominance of *religious* events at this specific life stage and is a finding relatable to notable changes in the belief system happening at this point in life. In addition, it may explain the observed decrease of religious events over the text course since descriptions of adolescence mainly fall into the first two to three bins which can be expected to deal with this part of live.

9.2.3 Polarity of life event domains

Having related both sentiment polarity and life event domain distributions to life stages, the only task combination left is the one between them. To this end, Figure 9.10 shows differences between the three scales of Moravian memoirs under investigation. In general, both manual and automatic annotations of the smaller subsample are once again very similar to each other and thus validate an upscaling. In the case of the *birth* of the subject, this means that it is almost exclusively depicted as neutral as possible - i.e. in the form of a factual report. This fits the assumed overall structure of Moravian memoirs where the notion of birth is expected to fall in the introductionary section dedicated to demographic facts (see Table 4.1). As somewhat expectable from the previously reported findings, the polarity distribution is different in the case of the *death* of the subject which is, in fact, one of the most positively depicted event domains in these sources.

The domains *education/work*, *housing*, and *no or other* are depicted similar to family-related ones and the overall polarity. This means that they are mostly depicted with a positive polarity albeit that a neutral one is not uncommon whereas a negative one is rather rare. This is even more true for *leisure* and *religious* events with the former being at least as positively depicted as the eventual death and the latter being only slightly less positive in the manual annotations. The positivity of leisure activities is not really surprising while the one of religious events could relate to the claim of Mettele (2007, 23) that an "estrangement from the [religious] community" was the only topical "taboo" for Moravian life narratives.

As discussed earlier, Moravian beliefs and the topical choices arising from them also explain why the *health* domain is, as a matter of fact, the one and only outlier polarity-wise in all three investigated scales. More specifically, there are a few positively connoted instances – most likely describing health *improvements* – but they do not fall in the interquartile range. Instead, it starts in

the neutral range and spans over the whole negative one in the smaller sample. In the fully upscaled version, the lower quartile is a little less negative but the polarity median is still below –.25 and hence quite negative. As a result, this finding seems to contradict the assumption of Faull and McGuire (2022, 143) that "[d]iscourse relating to illness can often be both positive and negative but should average out to positive overall when Moravian-specific vocabulary and contextual word usage is accounted for" since the fine-tuned model met both conditions. However, health issues were shown to correlate with the positively framed death so that the interpretation may again depend on the perspective: When considering the discourse level of the narrative, there may be more positivity not because illnesses are depicted this way but due to other intertwined topics. Thus, it is actually possible that choir helpers composed the final version of the life narrative in a way that fits a conversion and/or salvation narrative. The possibility of such a framing is further supported by the fact that bins increasingly dealing with health and death are also characterized by a notable increase of non-autobiographic remarks of choir helpers.

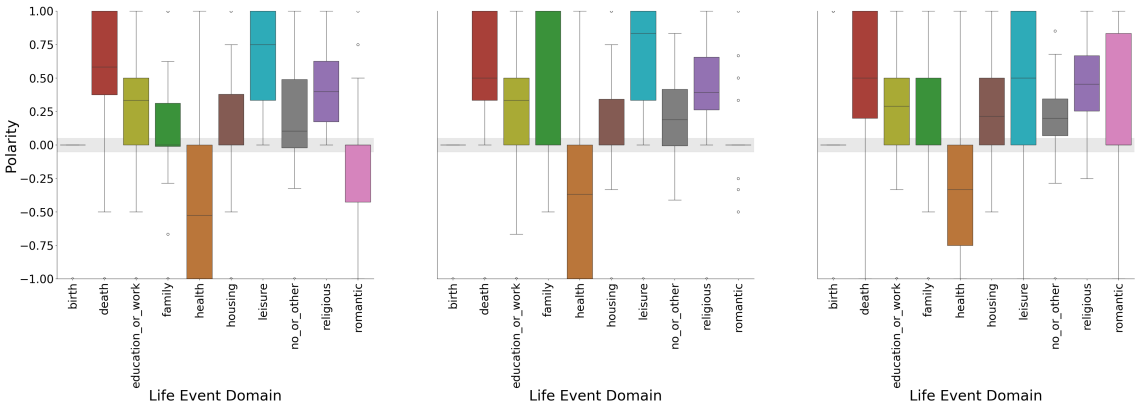

Figure 9.10: Sentiment polarity of life event domains in Moravian memoirs (36 manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right)

Methodologically, it should be mentioned that *family* events were attributed a higher polarity by gbert, while *romantic* events were predominantly annotated with a negative polarity manually and a neutral one by gbert, on the other. The former finding is less of an issue as the distribution of the full corpus becomes more similar to the smaller manually annotated one. In addition, it was similar to the one of other event domains so that the interpretation (= a rather positive depiction) stays consistent. By contrast, *romantic* events seemingly flipped the polarity when upscaled – albeit that still a few instances with a completely negative polarity remained. One possible reason is that the manually annotated sample was just not representative enough and hence more negative than one would assume in the case of romantic events. This potential mismatch could be related to the custom of drawing lots for matchmaking purposes in the 18th century community analyzed here (Becker-Cantarino, 2016, 185). Another explanation could be that the classifier has also learned to associate romantic events with a positive polarity when gaining "world knowledge" during its pre-training stage. This would imply that the amount of training data was just too low for a debiasing. Yet another potential resolution of the mismatch can be inferred from the respective confusion matrix (see Figure 8.2). It showed that gbert labeled a few family events as being romantic so that some of the positivity could actually rather relate to this other social event domain. This is why this specific finding is an apt one to be explored further with a close reading approach.

Close reading of romantic events in Moravian memoirs

In order to assess whether the observed polarity flip is either related to an (unwanted) artifact of imperfect classifications or actual differences between the analyzed samples, all 61 manually annotated instances of romantic events where compared to the 41 automatically annotated ones in the same and the 86 in the larger sample in a close reading setting. In the case of the smaller sample which was manually annotated as well, gbert missed six descriptions of arranged weddings,

five of break-ups, six of partner losses, and two of proposals. As a matter of fact, most of these false negatives are rather negative by nature for the respective experiencer which would fit a bias learned from pre-training data of gbert. Nonetheless, the model still found two cases of an arranged wedding and eight losses. One of the latter cases is Example (9.1) which actually deals with a family member instead of the biographical subject even though this is not as easily inferable from the respective sentence:

(9.1) *Sie ist auch bald darauf selig verschieden.*
('She passed away blessedly soon after.')

Therefore, it is difficult to actually assess a bias towards positively connoted romantic events. This is even more true when considering that some of the false negatives are quite implicit, such as in the case of the break-up described in Example (9.2). The few romantic events framed as positively as Example (9.3), by contrast, are more explicit and labeled correctly.

(9.2) *Darüber wurde sie böse und ging davon.*
('She became angry about this and left.')

(9.3) *Daher beschloß ich bald, zu heuraden, welches auch geschah 1737 und wurde den 23. Dezember getraut mit einer lieben {NAME}, geborene {NAME}in, die ich von Jugend auf liebhatte.*
('Therefore, I soon decided to merry, which happened in 1737, and on December 23rd I was married to my dear {NAME}, née {NAME}, whom I had loved since my youth.')

When disregarding false negatives, many romantic events were, as a matter of fact, depicted in the form of factual reports no matter if they are positive for the respective experiencer, as in the case of Example (9.4), or not, as in Example (9.5).

(9.4) *Die Trauung erfolgte am 16. Aug. des gedachten Jahres 1770.*
('The wedding took place on August 16th of the year 1770.')

(9.5) *1789 wurde sie Witwe;*
('She became a widow in 1789;')

Hence, one could conclude that the negativity found in the smaller corpus may actually rather be due to a sampling bias than one of the classifiers involved. In addition, it is worth highlighting that if there were such a disparity between more implicit negative framings and more explicit positive ones, that would definitely be another interesting content-related finding. However, it is up to future research with a larger sample size to prove that this is really the case.

Example (9.3) was also a training dataset instance which could offer an alternative interpretation of its true positive status. Still, the example is noteworthy because it was the only one of the class `romantic` which comprised the non-standard form *heuraden* of *heiraten* 'to marry'. It deviates from the standard due to a different diphthong in the first syllable and different onset voicing in the last one. These differences are regiolectal in nature but there is actually a more common graphematic variant in the latter case – namely *th* instead of *t* – according to Roth (2025, 102). Yet, this variant occurred only once as *verheirathete* 'married' in the training samples of romantic events. It is therefore quite remarkable that gbert seemingly learned more (combinations) of these variants when considering relevance scores of Transformer-Explainability for sentences with the subword tokens `##rath#ete`.

For instance, the first text in Figure 9.11 is basically the training sample when neglecting differences due to the manual cleaning of sentence boundaries not conducted in the case of the larger sample. The relevance scores underline that gbert indeed focused on the token *verheirathete* 'married' when assigning the correct label `romantic`. Similarly, the exact same token was the most relevant one in the case of the second text. However, the classification was in fact wrong as (i) it was another family member (= the father) who married and (ii) the subject traveled around in the aftermath. Hence, the sentence comprises two events and this is reflected in the confidence

text	label	birth	death	education_or_work	family	health	housing	leisure	no_or_other	religious	romantic
Br. ergriff gleich drauf die Orgel-Bauer-Profession, blieb in Dresden, u. verheirathete sich.	romantic	0.70%	1.54%	0.73%	8.29%	0.84%	3.96%	0.84%	1.65%	0.90%	80.55%
Weil aber mein Vater in seinem 80ten Jahre sich wieder verheirathete, so verließ ich Zürch 1737 im Herbst, ging nach Tübingen, und von da um Ostern 1738 nach Hamburg, wo ich ein Jahr blieb.	romantic	1.38%	0.88%	1.00%	15.53%	0.77%	31.23%	1.18%	0.73%	0.35%	46.94%
A. 1728. d. 28tn Nov. heyrathete ich meinen seligen Mann;	romantic	0.95%	7.65%	0.89%	8.24%	1.11%	1.84%	1.34%	17.36%	20.89%	39.72%
1739. heurathete er die nunmehrige Witwe u. sezte sich auf ein Stück land 3. Meilen von Bethlehem an der Lecha.	romantic	1.07%	2.72%	0.59%	7.41%	0.67%	3.03%	0.82%	1.17%	1.81%	80.71%

Figure 9.11: Transformer-Explainability based feature relevance table for romantic events with the subword tokens ##rath+##ete as predicted by a fine-tuned gbert (lighter colors correspond to less decisive tokens)

scores of .16 for `family` and .31 for `housing` which are lower than .47 for `romantic` but still higher than average (= .10). In this sense, this specific classification behavior is actually at least comprehensible and suggests that the model may be better than implied by the raw F1 score.

In the case of the third sentence, the model was less confident in its correct classification even though `##rath##ete` was again the most relevant subword combination which is remarkable as it has not seen the word *heyrathete* with a *y* in training instances. The same is true for *heurathete* which was seemingly the basis of the very high confidence for the correct label in the fourth and last text. In conclusion, these additional findings suggest that the model has implicitly learned these graphematic and regiolectal variants so that it is in fact adapted to Moravian sources of the 18th century. This can be counted as a major learning success.

Distributions over narrative time

As a final change in perspective, one could also consider polarity trajectories per life event domain over the narrative albeit that this is not worth the effort for the neutrally depicted *birth* of the subject. The remaining nine trajectories are shown in Figure 9.12 and illustrate that the polarity of most domains actually varied over the text course.[24]

In the case of the *death*, it is worth highlighting that prolepses at the beginning of the narrative are less positive than ones in the end whereas ones in the middle are even negative. When assuming that death accounts span the second half of the narrative, the mention in the middle can be seen as the beginning of this topical line. Hence, it is not unlikely that the death itself is depicted as tragic and hence negative but subsequently characterized as more positive than it actually is – likely in order to follow the salvation narrative. This is not necessary when describing events from the *education or work* domains which is probably the reason why their polarity trajectory basically meanders around the rather positive mean value. The same can be said about *romantic* events although, for some reason, the few occurring in the second last bin are depicted neutrally or even in a negative way.

There is also a negative low at the end of the first third of the narrative in the case of *family* events which is followed by a high and an overall polarity rise. A similar rise is observable for relevant *health* events albeit that this is again the only domain which stays almost exclusively in the negative range. The only exceptions occur in the beginning and the end. When considering that the beginning section is the most likely one to deal with children and adolescents, it is possible to relate the first high with overcoming the risks of the high infant mortality rate and hence actual health improvements. The subsequent rise, by contrast, is mostly likely again related to the salvation narrative. To a lesser degree, this may also affect *housing*, *no or other*, and *religious* events which all involve an overall rising polarity pattern. This pattern is otherwise only lacking in the case of the few mentioned *leisure* events which have a negative low when described in the first bin and a second (more neutral) at their latest occurrence in narrative time. However, due to their scarcity, it is difficult to motivate this deviation from the overall pattern.

[24] A further rescaling to the narrated life stages does not lead to additional insights (see Figure B.1).

Last not least, it is worth highlighting that the trajectory of the residual class `no_or_other` mirrors the overall pattern (see Figure 9.2) the most. This can be seen as one further confirmation for the fact that choir helpers aimed at composing a narrative of this kind because one would actually rather expect such (non) events to also meander around the mean sentiment polarity.

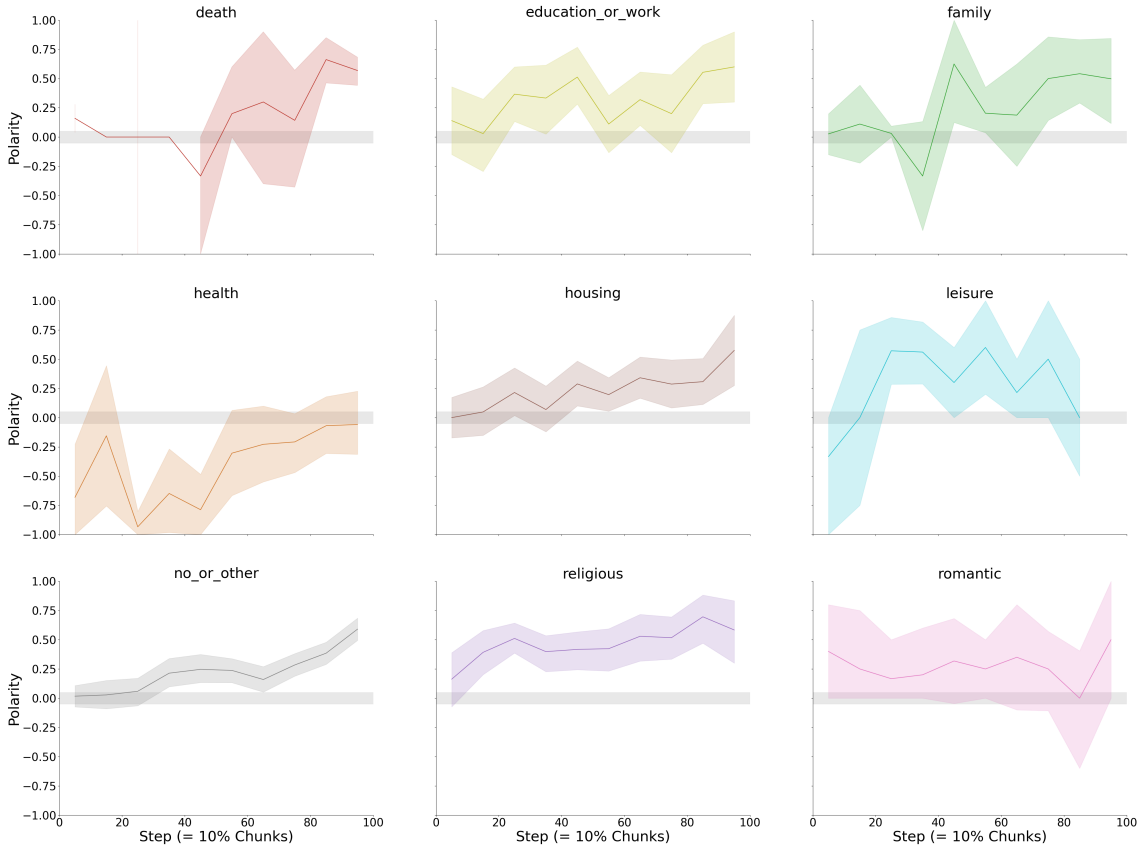

Figure 9.12: Sentiment polarity of life event domains in 89 Moravian memoirs over narrative time as predicted by fine-tuned gbert models

9.2.4 Summary

Methodologically, this first content analysis has shown that fine-tuned gbert models are in fact apt for an upscaling from a manually annotated subsample to a larger sample. More specifically, most observable patterns were similar in the case of the smaller sample no matter if analyzed based on manual or automatic annotations despite an F1 score of .81 for a sentiment (see Section 6.3), .49 for a life stage (see Section 7.3), and .71 for a life event domain classification (see Section 8.3).

In a similar vein, there were fewer differences between the smaller sample and the larger one as one could assume implying that the former was rather representative for the latter. By extension, this can already be seen as a confirmation for the hypothesized general pattern Moravian memoirs follow – especially when considering **sentiment polarity**: More specifically, analyses of polarity trajectories revealed considerable differences between narrative and narrated time with the former matching hypotheses closer than the latter (see Figure 9.8). For instance, a childhood peak was only observable when relating manual sentiment to life stage labels and hence rescaling to the narrated time. Yet, the peak diminished in the case of automatic annotations although this may be a side effect of the imperfect classification. The stark difference in the depiction of the end of the narrative (= more positive) and narrated time (= less positive) is, however, most probably no such effect and hence worth investigating further.

One first endeavor to this end was a closer look at **life event domains**. It hinted at the possibility that memoirs were composed in a way that fits a conversion and/or salvation narrative no matter the actual depiction of life stages and event. This may also explain why memoirs tend to mostly follow

the life course in chronological order while still comprising stylistic devises like pro- and analepses which break that order (see Figure 9.4). In addition, the position of non-autobiographic remarks in the beginning as well as the final third of the narrative may also contribute to these findings. Similarly, the status of Moravian memoirs as a rite of passage tied to a funeral setting coincides with fitting topical foci, such as (lethal) illnesses and the eventual death – especially in the last third of the narrative (see Figure 9.6). The religious events mentioned in these sources predominate during adolescence (see Figure 9.9) and hence the first third of the narrative. In summary, this implies that these sources indeed fit their role of depicting the most important events of their time and belief system in a somewhat patterned narrative while leaving room for individual differences on the narrated side.

9.3 Gender differences in Moravian memoirs

Women's lives can be expected to have differed from men's in the late 18th century in a way that could be reflected in biographical sources. In the broadest sense, this can already be inferred from the fact that only 4 % of the ADB biographies about individuals who lived between 1750 and 1850 are dedicated to women (see Section 4.3). Since this gender gap is somewhat typical, Faull (1992, 24) chose to focus on women when editing Moravian memoirs – even more so because there are hardly any other ego-documents of female artisans and peasants from that time. This is why the situation is more balanced in the case of these sources so that a fully balanced subcorpus for manual annotations could be compiled. The full corpus, by contrast, has actually a slight female bias with almost 61 % of the memoirs describing lives of girls and women (see Table 5.1).

Related work yielded mixed results when trying to identify gender differences in Moravian memoirs: Mettele (2007, 33), for instance, claimed that "gender differences are fairly negligible" due to Moravian beliefs whereas Faull and McGuire (2022, 144) found more positive emotions in women's memoirs than in men's. As the latter study is based on a computational sentiment analysis – albeit with the different methodological approach of using a custom dictionary –, Section 9.3.1 also starts with sentiment polarity. However, unlike in the previous section, narrative and narrated time will be discussed jointly. Section 9.3.2 deals with possible differences in life stage distributions before Section 9.3.3 relates labels from all three classification tasks of this thesis to each other. Section 9.3.4 then summarizes the most import findings before Section 9.4 adds another dataset in order to look at potential genre differences instead of a demographic factor, such as gender.

9.3.1 Gender differences in sentiment polarity

When considering the text level, memoirs of female Moravians are described in a more positive way than the ones of their male counterparts in all three samples (see Figure 9.13). However, the samples still differ slightly but notably from each other. For instance, the mean polarity of the 18 manually annotated texts with male subjects is with .08 almost in the neutral range. This changes when scaling up to the full corpus where the interquartile range is almost completely in the positive range and the mean increases to .15 although it is still below the female one (= .27). In this sense, the observed gender-based differences in overall polarity are in line with the findings of Faull and McGuire (2022, 144) even though different methods were used for the sentiment annotation (= a fine-tuned BERT model instead of a sentiment dictionary) and larger samples (= 36 or 89 instead of 21) in another writing languages (= German instead of English) were investigated.

Observing gender differences in sentiment polarity with different systems

Due to the availability of various sentiment classification systems, we previously investigated to what extend gender differences in overall text polarity can be inferred with different systems (Brookshire and Reiter, 2024, 96–97). In the paper, we focused on a few selected systems but further ones are considered here. More specifically, the six most performant systems per system type (see Section 6.3) are compared to each other in Figure 9.14 with respect to the smaller manually annotated sample. As a matter of fact, annotations of most systems lead to the aforementioned

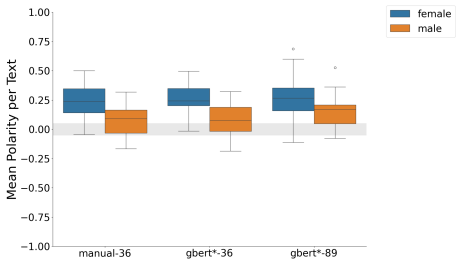

Figure 9.13: Distribution of sentiment polarity in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right) grouped by subject gender

conclusion that texts with a female subject are depicted more positively than their male counterparts. This is noteworthy as the systems vary considerably with respect to F1 scores (see Table 9.2). so that one can come to the conclusion that the gender difference is so apparent that even mediocre performances are enough to detect it. Although it is obviously better to use the best system possible, this finding has implications for settings where options for model optimizations are limited since a system comparison could be a viable option in these cases (see Elkins, 2022, 46).

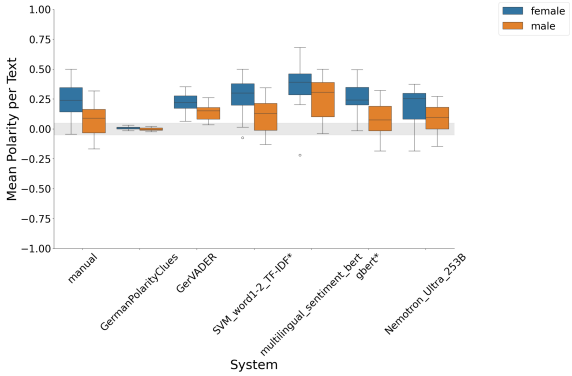

Figure 9.14: Mean polarity distributions of 36 Moravian memoirs grouped by subject gender when annotated with selected systems (systems fine-tuned/trained on Moravian data marked with *)

Still, there are a few differences between the systems under consideration which are worth underlining. The most obvious outlier is the dictionary GermanPolarityClues with its strong preference towards predicting the label `neutral`. This is most likely related to out-of-vocabulary issues as suggested by the coverage analysis (see Table 6.3). However, it should also be noted that GerVADER covers even less tokens and lemma forms despite being closer to the manual annotations so that the usefulness of this kind of analysis for an a priori assessment is somewhat diminished. The main difference between both dictionary-based approaches is that the latter adds contextual rules which seemingly does not only increase basic performance scores but also the applicability to downstream tasks. In this sense, the finding supports related work which showed that such an addition is the most promising one in the case of the widely used dictionary-based approaches towards sentiment analysis (see Section 3.1.2).

Another way to compare the results of the different systems is to look at effect sizes measured with Cohen's D (see Table 9.2). In this context, it is noteworthy that the GerVADER annotations are actually closer to the manual ones than in the case of some better performing systems. More specifically, annotations from both supervised learning approaches (= the most performant SVM setting and the fine-tuned gbert) also suggest a large effect as they exceed the common threshold of .8 but lead to an overestimation. Conversely, the effect is underestimated to a medium effect when based on annotations of the most performant zero-shot approach (= the largest LLM considered). An even lower medium effect would be attested based on predictions of the off-the-shelf dictionary and BERT model. The decrease in effect size for the last three systems is in line with the respective F1 scores so that the actual effect is most likely a large one.

System	Type	Cohen's D	F1 score
manual	–	1.08	–
GermanPolarityClues	dictionary	.69	.47
GerVADER	dictionary + rules	**1.06**	.58
SVM_word1-2_TF-IDF*	SVM*	1.13	.70
multilingual_sentiment_bert	BERT	.58	.42
gbert*	BERT*	1.21	**.81**
Nemotron_Ultra_253B	zero-shot (LLM)	.75	.77

Table 9.2: Effect size measured with Cohen's D of mean predicted polarity on gender metadata in 36 Moravian memoirs by selected systems (value closest to manual annotations in bold, systems fine-tuned/trained on Moravian data marked with *)

Distributions over narrative and narrated time

Polarity differences are not restricted to the overall text level as illustrated by Figure 9.15. In the case of the **narrative time**, this applies first and foremost to the second and to a lesser degree also the seventh step. The differences between the manual and automatic annotations are again minor whereas the upscaling of the latter to the full sample fits expectations from the overall distributions. More specifically, while the second step forms a high in texts about girls or women, it forms a low in texts about boys or men which is actually in the negative range in the case of the smaller sample. However, similar to the higher overall polarity found in men's memoirs, this low falls into the neutral range after upscaling. Similarly, one can observe a second low at the seventh step in men's memoirs which is mostly neutral in the smaller sample and rather positive in the larger one. Said low contrasts the overall rising found in the case of the opposite gender in both samples.

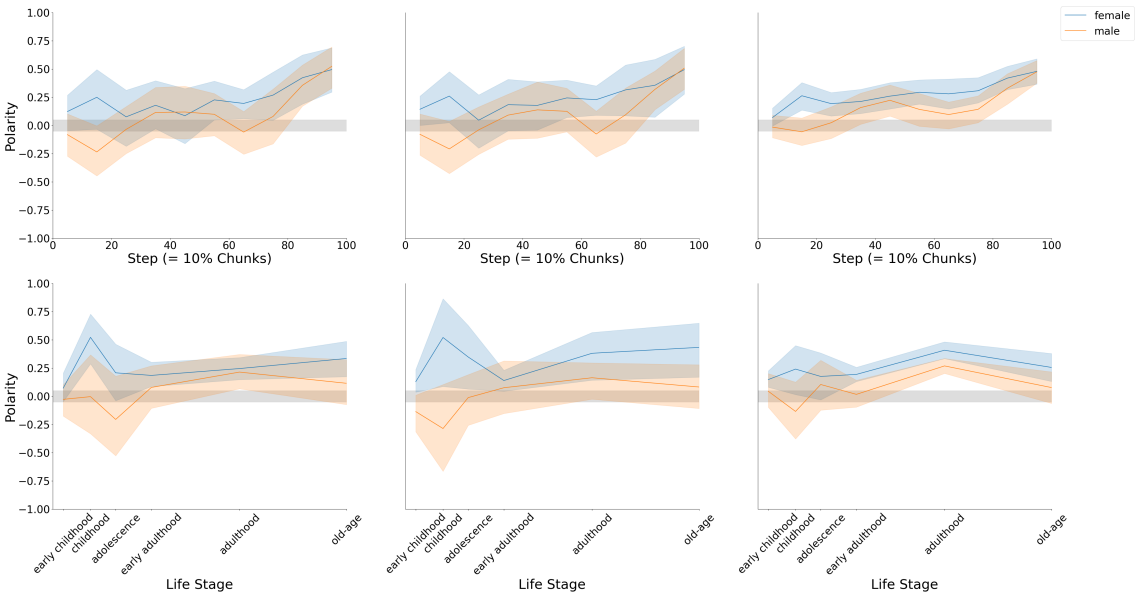

Figure 9.15: Comparison of Moravian sentiment trajectories (36 memoirs manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right) over narrative (top) and narrated time (bottom) grouped by subject gender

After rescaling to the **narrated time**, it becomes apparent that the first female high concerns childhood experiences whereas the male low rather concerns adolescence in the case of the manual annotations while also being related to childhood in both samples when labeled automatically. As a result, a high during male adolescence seems to emerge which would not only be a stark contrast to the female pattern but also to the expected "fall through adolescence" (McGuire, 2021, 133). Given

that the classifier tends to confuse *adolescence* with *early adulthood* and that a second male low emerges during the later life stage when upscaled, it is, however, up to future research to further validate the finding.

The same is true for the relationship between the depiction of adulthood and old-age in women's memoirs because both life stages are consistently in the positive range albeit that a falling pattern emerges during upscaling. It is similar to the one found in men's memoirs and remarkable because it contrasts the narrative pattern observed in Section 9.2.2 where gender differences where not taken into account. It also contradicts expectations towards a conversion/salvation narrative so that the polarity trajectory over narrative time is again closer to the expected one. Interestingly, this applies more to female subjects than male ones as shown in Figure 9.16.

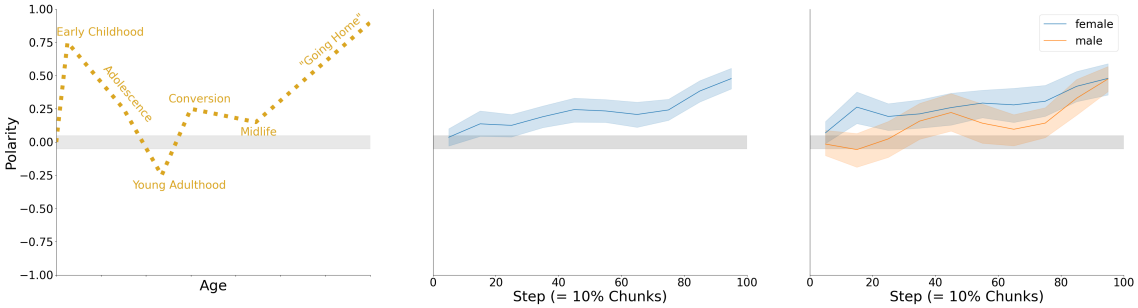

Figure 9.16: Comparison of Moravian sentiment trajectories (idealized based on McGuire (2021, 133) left, 89 memoirs predicted by a fine-tuned gbert left and grouped by subject gender right)

The expected *childhood high* emerged only slightly during upscaling when neglecting gender differences although it is observable in all female samples under consideration. As the larger sample comprises more texts about girls or women, one could come to the conclusion that it may have been leveled off before due to a male childhood low. As a matter of fact, McGuire (2021, 135) also observed a lower polarity in the beginning of 36 English men's memoirs from the British Fulneck compared to 86 women's from the same place. Faull and McGuire (2022, 143) found the same relation between gender and polarity in the first third of the narrative in their smaller British sample from Fatter Lane albeit that they identified an opposing one with respect to English memoirs from Bethlehem. These mixed findings in related work highlight potential individual and/or region-specific differences. Hence, it is remarkable that the expected childhood high was not found in these previously analyzed corpora and only found as a *girlhood* high here. However, it should be added that Faull and McGuire (2022, 144) attributed some of the negativity they observed with their "current sentiment dictionary and tagging limitations" so that it is possible that a more similar pattern emerges in their future studies.

With respect to the subsequent trajectory, it is worth mentioning that Faull and McGuire (2022, 143) also identified an overall rise in polarity in women's memoirs from Bethlehem whereas the two previously analyzed British corpora included a more notable "midlife low" so that this may also be a regional peculiarity. This is even more true as said expected low was visible in all corpora as well as the current one in the case of the male polarity trajectory. Yet, the chronological rescaling does not suggest that said low falls into *midlife* as there is actually an *adulthood high* observable in the case of the German memoirs under investigation in this work. This is why this specific disparity will be investigated further in Section 9.3.2.

The observed gender differences are also relatable to the basic emotional arcs identified by Reagan et al. (2016, 5) in the sense that men's memoirs seem to follow the "Man in a Hole" pattern and women's memoirs the "Cinderella" story. Brown and Tu (2020, 267–269) called the latter trajectory "Quest" and their "harmonic expansion" reveals another typological difference: Due to the flattened second peak in (auto)biographies with female subjects, one could argue that they are a little less complex or "harmonic" polarity-wise than their male counterparts (see Figure 3.5). In addition, it is worth highlighting that the harmonicity seems to decrease during upscaling (see Figure 9.15). However, after the chronological rescaling to the narrated time, men's memoirs become more similar to what Reagan et al. (2016, 5) called "Oedipus" and women's rather follow

the "Icarus" pattern with Brown and Tu (2020, 267) calling the same trajectories "Tragedy" and "Loser's tale" without gender-based oscillation differences. Therefore, it can be stated that the two versions of the underlying life narratives in fact lead to different conclusions when just considering these typologies.

In summary, sentiment polarity distributions on the text level already show that memoirs about female subjects of Moravian memoirs tend to be more positive in general which applies to trends over narrative and narrated time as well. The polarity difference is most pronounced in the beginning of the narrative and may be even more prominent with regard to childhood experiences. In this sense, it cannot be ruled out that the hypothesized pattern might include a female bias in the beginning and a male one at the end of the second third of the narrative (= the alleged midlife section). It should, however, also be noted that the female bias contrasts related work where women's memoirs matched the expected patterns less closely.

9.3.2 Gender differences in life stage distributions

Possible gender-based differences with respect to earlier life stages are also inferable from the relative amount of sentences dedicated to each life stage. As illustrated by Figure 9.17, memoirs with female subjects comprise more information about *early childhood* than their male counterparts. By contrast, there is a broader interquartile range in the case of male *childhood* in the smaller sample which approximates the early childhood distribution when upscaled to the larger sample. In a similar vein, the other distributions suggest that women's memoirs focus slightly more on *early adulthood* whereas men's memoirs describe the later stages *adulthood* and *old-age* in more detail.

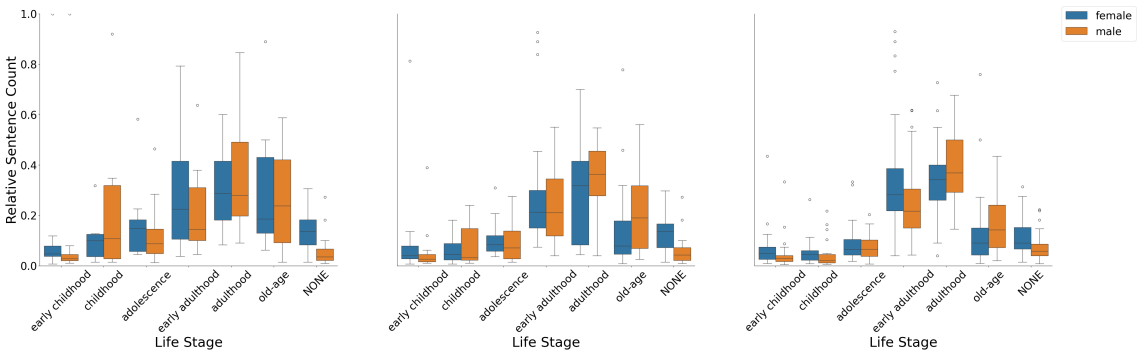

Figure 9.17: Distribution of life stages in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right) grouped by subject gender

These disparities are noteworthy because they appear to be somewhat independent from the respective ages at death (see Table 9.3). For instance, the fact that one girl and one boy died during early childhood is reflected by outliers in the overall distributions. However, this is only also true in the case of the manual annotations for the sole (male) individual who died during childhood. In a similar vein, the corpus comprises only women who died during early adulthood. Yet, respective memoirs account only for a very high whisker in the case of the manual annotations and some of the outliers in the case of the automatic ones.

Considering individuals who died during adulthood, it is worth mentioning that these were mostly female even though men's memoirs include slightly more passages about said life stage. The same is true for old-age in the larger sample. However, one should also mention that it seems like the classifier confused the labels `early adulthood` and `old-age` with `adulthood` in the case of women and vice versa in the case of men. Thus, some of the possible inferences should be taken with a grain of salt and further validated in future work based on larger samples. Still, the interesting conclusion that earlier stages may be more prominent in memoirs with female subjects and whereas men's memoirs focus on later ones is not completely unlikely given the distributions.

Another interesting and less debatable finding is that memoirs about girls and women comprise more *non-biographical remarks*. As a result, this means that their memoirs in fact deal less about their lives than in the case of male Moravians. It is unclear if this disparity is a special feature of

the selected German sample from Bethlehem because – to the best of my knowledge – no other study has so far looked similarly into life stages in general and non-autobiographical passages of Moravian memoirs in particular. The only prior work which could be related to this finding is the pronoun usage comparison of Faull and McGuire (2022, 141). It showed that Moravian women from Bethlehem used male pronouns approximately twice as much as female ones whereas women from Fetter Lane used both grammatical genera with almost equal frequency. Even though these analyses are coarse-grained in the sense that gendered pronouns could also refer to other people, the result that women appear to talk less about female individuals like themselves is in line with the main finding of the life stage classification task, here.

Life Stage	**Subsample**		**Full Sample**	
	Females	**Males**	**Females**	**Males**
Early Childhood	1	1	1	1
Childhood	–	1	–	1
Adolescence	–	–	–	–
Early Adulthood	1	–	3	–
Adulthood	7	3	15	6
Old-Age	8	12	34	25
Unclear	1	1	1	2
Total	18	18	54	35

Table 9.3: Life stage at death in Moravian subsamples

Distributions over narrative time

When looking at trends over the text course (see Figure 9.18), most findings prevail albeit that some new peculiarities arise. With respect to *early childhood* experiences, it was previously mentioned that the sole individual who already died during this early stage was a boy. It is thus not surprising that all male bins of the manually annotated subsample include sections about this stage of life. It is, however, more surprising that (i) this also applies to female bins and (ii) only to female bins in the case of *childhood* even though all samples are balanced with respect to childhood deaths. Hence, this finding is in line with the aforementioned observation that women wrote more about their (early) childhood experiences. This has implications for the discussion of whether there is only a (female) childhood high, as one would expect to find one in the case of both genders.

In addition, it is worth highlighting that the classifier had issues identifying early childhood, childhood, and adolescence when mentioned after the first third of the narrative in the case of female subjects whereas it mostly only had issues in the last two bins in the case of male ones. This gender disparity is noteworthy but may be related to life event domains. Furthermore, it has some implications for the fact that the gender differences of the first three life stages after the first third of narrative time somewhat level off when upscaled. More specifically, it could be the case that the gender difference of female subjects comprising more information about earlier life stages would be even more pronounced if the classifier did not have said bias. In a similar vein, the aforementioned classifier biases of confusing the labels `early adulthood` and `old-age` with `adulthood` more often in the case of women and vice versa in the case of men affect the interpretation of later life stages and later narrative passages. More specifically, it seems as if most differences concerning the proportions of (early) adulthood level off when considering automatic annotations. However, the smaller manually annotated sample suggests that men described early adulthood experiences more briefly and mostly only in the first third of the narrative before covering later stages of life whereas women's memoirs show the opposite pattern.

It is noteworthy that *non-(auto)biographical* sections are more dispersed over female life courses than over male ones. They appear to have the same bracketing purpose in the case of both genders but the increase over the second half of the narrative is more pronounced in the female case.

By contrast, the fewer non-biographical remarks in memoirs about boys and men mostly occur only at the first and the last two bins. In addition, there are some in the middle but they are more scattered depending on the sample under investigation. Still, they seem to again co-occur to a certain degree with descriptions of *old-age* which men's memoirs covered in more detail while women's memoirs comprised more non-biographical remarks. It is thus probable that the comparably high proportion of the latter in the beginning are most probably introductory in nature. Later ones, by contrast, more likely serve the purpose of balancing out the more negative sides of illnesses and death in order to fit the salvation narrative as hypothesized earlier.

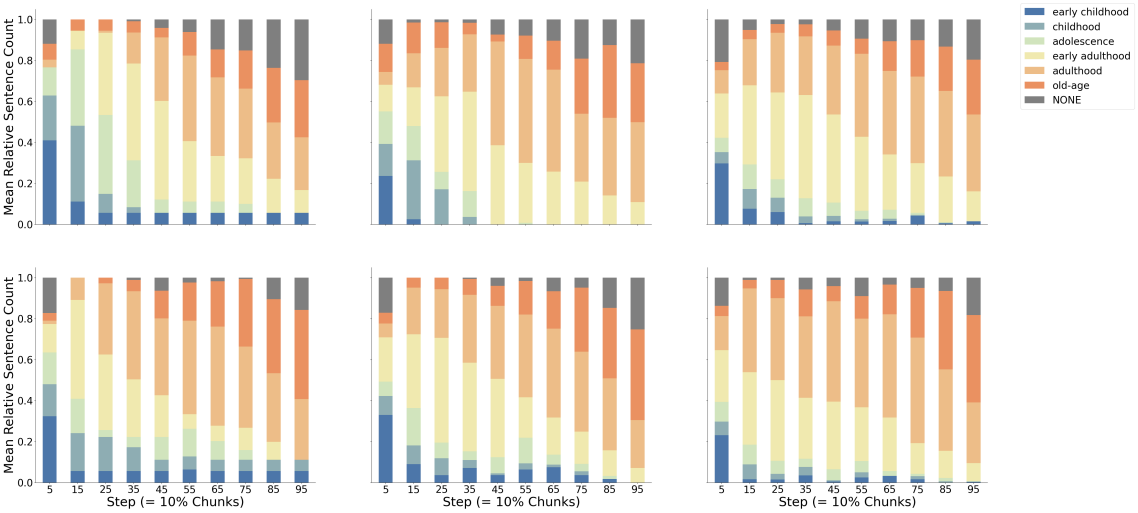

Figure 9.18: Distribution of life stages in Moravian memoirs over narrative time (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right) grouped by subject gender (female top, male bottom)

Finally, it is worth highlighting, that women's memoirs are characterized by a larger proportion of non-biographic remarks as well as early adulthood experiences at the end of the second third of the narrative in all samples under consideration. Given the assumed balancing purpose of the former, this may in fact explain the lack of a female midlife low from the previous subsection. Nevertheless, this interpretation also implies that women either experienced fewer negative aspects of midlife than men and/or wrote less about them.

9.3.3 Gender differences in life event domain distributions and their polarity

The overall distribution of most life event domains is quite similar no matter the subject's gender as shown in Figure 9.19 even though there are a few exceptions: Most notably, the domain *education or work* seems to be the most gender-specific one as its distribution of sentences per text is more skewed towards men's memoirs than any other domain towards one gender or the other. By contrast, *health* events appear to be slightly more prominent in memoirs with a female subject. However, similar to *family* events, which are only minimally more associated with female Moravians on the overall text level, the classifier overestimated the gender-specifity so that both are better deemed gender-neutral. This is also plausible given that (i) the former are among the primary topical foci in these sources and (ii) there is no obvious reason why life events of family members of girls or women are more likely to be described than ones of boys or men due to the choir system (see Section 4.1.1).

The slight associations between *housing* changes, *leisure* activities, and *no or other* events with the male gender are more difficult to interpret. This is partly due to the fact that especially the latter domain becomes less gender-specific when scaled up to the full sample so that sampling effects cannot be totally ruled out. Furthermore, it should not be forgotten that it is the most difficult event domain to interpret by definition. Similarly, *romantic* events may be slightly more common in women's memoirs but are too rare overall to rule out sampling issues.

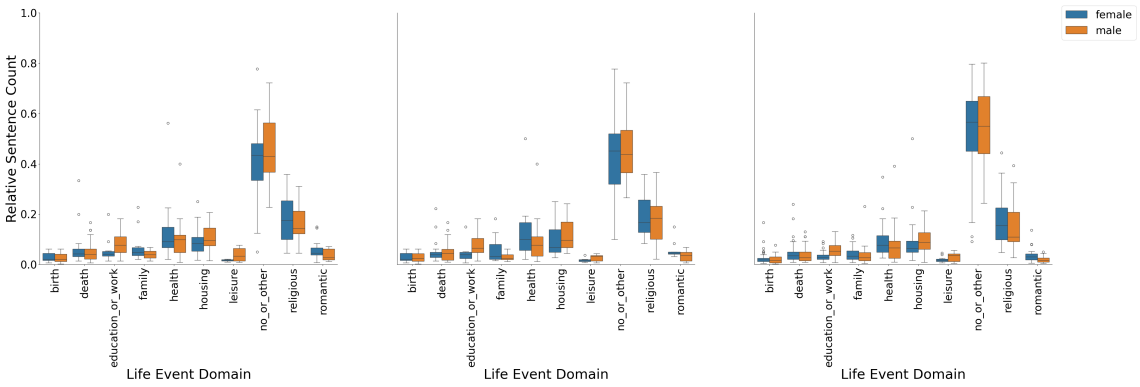

Figure 9.19: Distribution of life event domains in Moravian memoirs (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right) grouped by subject gender (female top, male bottom)

Distributions over narrative time

As inferable from the ungendered distributions over narrative time, the subject's *birth* is always only mentioned in the beginning while her or his *death* is more dispersed over the text course. However, Figure 9.20 illustrates that this last event is mentioned earlier in the case of female subjects even though this difference averages out after scaling up to the full corpus so that it may just be a peculiarity of the smaller sample. By contrast, the merged domain *education/work* is not only more prominent in the case of male subjects as already mentioned but also lacking in the end of texts about individuals of the opposite gender. When considering the final bin as some kind of wrap-up, this finding may imply that work accomplishments were more likely described in the case of men. In this sense, it is in fact another noteworthy disparity related to this event domain.

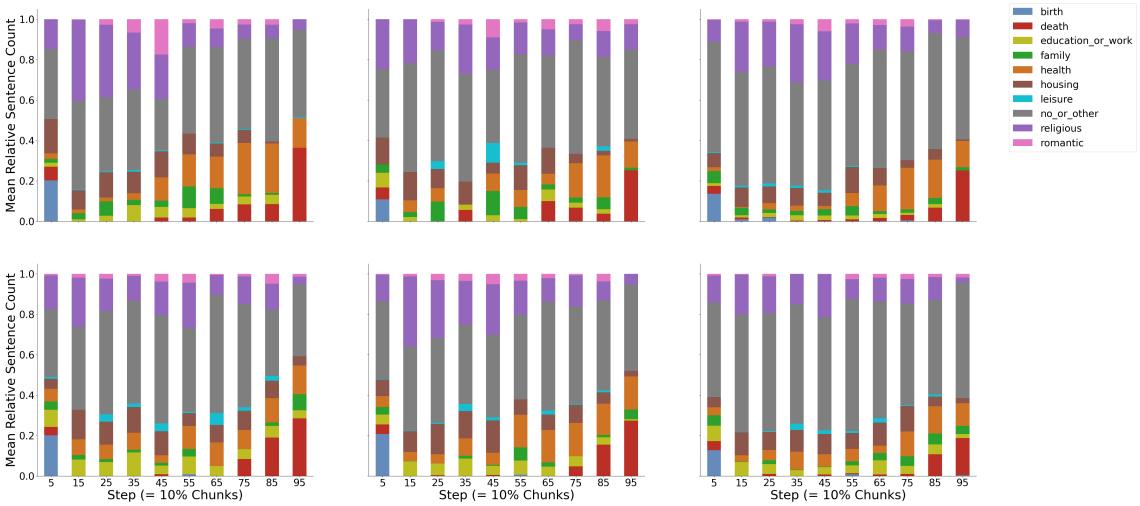

Figure 9.20: Distribution of life event domains in Moravian memoirs over narrative time (36 manually annotated left, the same predicted by a fine-tuned gbert middle and all 89 right) grouped by subject gender (female top, male bottom)

The distribution of *family* events over the text course is interesting because they only lack in the final bin of memoirs with female subjects while occupying slightly more text proportions in the middle of the narrative than in the case of male subjects. As a result, these two differences average out over the whole text course which explains why this event domain was seen as "minimally more associated with female Moravians" before. Given the proportions in the end of the narrative and the fact that this section was usually written by the choir helper, this finding suggests that Moravian men did not write as much about their family in their original autobiographical draft. The respective information was instead added later in order to make the final account as total as possible.

Considering *health events*, it is noteworthy that the proportions increase over narrative time in women's memoirs in which they are slightly more common while being more dispersed in the case of male subjects. As it is quite unlikely that boys or men encountered health issues earlier in their lives than girls or women, one could come to the conclusion that these distributional differences are in fact related to text composure and hence framing purposes. The same is most probably true for mentions of *housing* changes which decrease over narrative time in the case of female subjects while again being more dispersed in the case of male ones.

Leisure activities are more prominent in the case of men's memoirs but quite rare overall. This seems to be due to their lack in initial and closing sections. Interestingly, events of this kind are overestimated in the smaller sample by the classifier in women's memoirs so that they may be even more associated with the male gender than the overall distribution suggests.

With respect to the widespread mentions of *religious* events, it is worth highlighting that they occupy larger proportions in the first half of the text course of women's memoirs whereas there is one peak in each third in the case of the opposite gender. This may hint at age-related differences due to the overall chronological life description found in these sources. However, the aforementioned observations that (i) religious events are strongly associated with adolescence and (ii) men's memoirs comprise more descriptions of said life stage in later parts of the narrative also suggest that this may just be another kind of narrative framing. In a similar vein, the prominence of the *romantic* domain in the middle of women's memoirs may also be due to the different distributions of life stages over narrative time depending on the subject's gender.

In summary, one should again highlight that many domains are less dispersed in women's memoirs. As a result, it actually seems as if these texts are more structured based on topical foci than their male counterparts. It is therefore not surprising that they mostly focus on illnesses and the eventual death in the end as expected. In the case of men's memoirs, by contrast, this "going home" section is intertwined with more domains, such as education or work, family, and housing.

TF-IDF analysis of gender differences in the beginning of narrative time

One possibility to zoom in even further into the observed gender differences is to complement the insights from the classification tasks with more traditional approaches from the field of information retrieval, such as TF-IDF. In order to further investigate the observed gender differences in the beginning of the life narrative, this was operationalized by splitting sentences from the second bin into subcorpora based on the subject's gender. The results are summarized in Table 9.4 where a TF-IDF score of .9 is used as a threshold.

The TF-IDF values suggest that women's memoirs are characterized by topics like **religious initiation** (*getauft* 'baptized', *chor* 'choir'), **health** (*gesund* 'healthy', *krank* 'sick'), and **education** (*erziehung* 'education', *anstalt* 'institute', *erziehen* 'educate'). Men, by contrast, seem to rather focus on the topics **travelling** (*reisen* 'travel', *schiff* 'ship', *reiste* 'traveled', *fremde* 'foreign [place]') which is also supported by various place names and years, **work** (*arbeit* 'work', *arbeitete* 'worked', *profession* 'profession'), and **war** (*soldaten* 'soldiers', *hauptmann* 'captain'). Most of these findings are in line with the previously discussed result of the life event domain classification that women's memoirs focus more on religious events and men's on education or work.

One slight contradiction between both analytical approaches to the initial part of the narrative is that health events are actually more common in men's memoirs. It is, however, not far-fetched to relate them to experiences of war and therefore a slightly different vocabulary. Similarly, housing changes were found to be quite common no matter the gender with the classification task but may be grouped slightly differently when looking at top ranked TF-IDF words. More specifically, it seems as if the female corpus comprises more instances of housing changes related to the Moravian choir system whereas the male ones lean more towards migratory and military experiences.

Interestingly, some results are also in line with the life stage annotations. More specifically, the latter hinted at the peculiarity that women focus more on earlier and men on later life stages – even at this early point in the narrative – (see Figure 9.18). This difference can also be inferred from semantic fields, such as baptism and education, for girls and young women on the one hand and (military) work for young men on the other.

Value	Women's memoirs	Men's memoirs
> 5.	ihre	
> 4.	ihrem, gem	
> 3.		
> 2.		bein
> 1.	vergnügt, getauft, gesund, schw, balde, schwestern, ziehen, jugend, erziehung, anstalt, übrigen, solchen, derselben, tochter	erst, große, reisen, essen, arbeit, soldaten, schiff, eigenen, reiste, herbst, holland, böses, beim, unsre, mal, folgenden, zusammen, melchior, fremde, gerechtigkeit
> .9	pensilvanien, manchmal, verbrachte, suchen, 13ten, herrnhaag, gern, heilige, wahrem, o, hernach, stille, freund, rege, erbarmen, verlohren, muß, erlaubnis, jedoch, erziehen, krank, manchen, 10, verlor, angefangen, harte, hätten, brauchte, oncle, sünderin, seeligkeit, chor, seelig, dienste, herrschaft, wahre	jüngling, ums, hansen, arbeitete, geschehen, geistes, dresden, wunden, dadurch, spielen, ewigkeit, 1733, scheune, hauptmann, obgleich, basel, altona, 1729, sinn, gründlich, stellen, deswegen, wohnte, bösen, tröstete, dies, sobald, indianer, junger, folge, wohnten, vorher, acker, pferde, wunderbar, kurzer, profession, orte, lauter, schluß, gefängniß, faßte, alten, figuren, klarheit, hhuth, erhörte, zauchtel

Table 9.4: Most relevant tokens at step 15 in Moravian memoirs dedicated to 54 women and 35 men as identified with TF-IDF values

Distributions over narrated time

After rescaling to the narrated time, some of the assumed interrelations between life event domains and life stages become even more apparent (see Figure 9.21). For instance, the merged domain *education or work* is not only more prominent in texts about male individuals in general but also with respect to adolescence and old-age – at least in the smaller sample. In the former case, this is especially noteworthy due to the strong association of said life stage with religious events which is actually more prominent in the case of the opposite gender. Hence, this finding is in line with the aforementioned focus/dispersion theme that appears to be associated with the two genders.

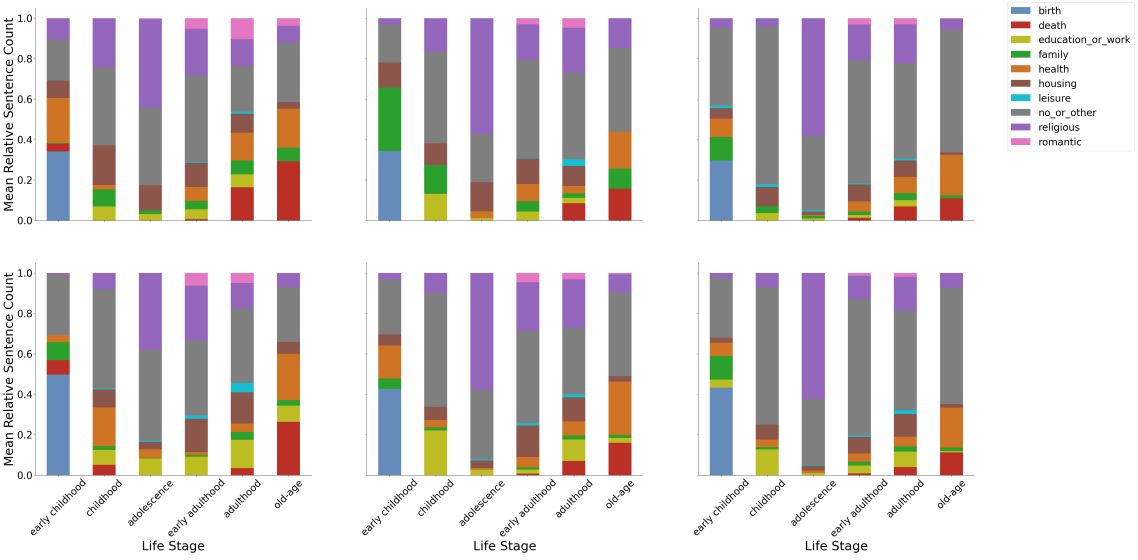

Figure 9.21: Distribution of life event domains in Moravian memoirs over narrated time (36 manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right) grouped by subject gender (female top, male bottom)

The general trend is seemingly more on the narrative side as *family* events are in fact more dispersed over the lives of girls or women than their male counterparts – even when disregarding the overestimation of the classifier. In the case of the manual annotations, it is actually remarkable that such events are not described during early girlhood while this life stage appears to be most heavily associated with family events for the opposite gender. This disparity is difficult to motivate based on real world differences as experiences of family members are by definition independent from the gender of the biographical subject. Hence, it appears to be the case that men deemed events of family members more relevant when they occurred during their early childhood whereas women deemed them more important during later life stages.

At first glance, it looks like there are also gender-related differences with regard to the life stages during which *health* issues are mentioned. However, these are mostly in line with the distribution of the subject's *death*. In addition, the classifier underestimated health events in early girlhood while overestimating ones during early boyhood in the smaller sample. Given that all these differences average out when scaled up to the larger sample, it is quite likely that the observable disparity is only related to peculiarities of the smaller sample. The same reasoning may explain why the female decrease and male increase of *housing* changes over the life course is only found in this sample.

Leisure events are a possible candidate for a gendered domain as they are most commonly found in descriptions of male adulthood – even more so when acknowledging that the classifier overestimated occurrences during female adulthood. In a similar but opposing manner, *romantic* events of female subjects are underestimated during said stage of life so that this event domain may actually be even more associated with this gender than implied by the overall text proportions.

Another remarkable finding is that *religious* events did in fact occur later in lives of Moravian men than in women's lives – at least in the manually annotated subsample. Thus, the aforementioned disparity with respect to the narrative time may actually not only be "another kind of narrative framing" but also a result of real world differences. Interestingly, the disparity becomes less obvious in the larger sample so that it may also be a peculiarity of the memoirs under consideration and thus up to future upscalings to interpret this finding.

Overall polarity of life event domains

When comparing life event domains with respect to their polarity, the first notable observation is that distributions vary more between the three samples than in the ungendered case (compare Figures 9.10 and 9.22). This appears to be mostly related to the slight positivity bias of the fine-tuned gbert model which affects different combinations of life event domains and gender of the biographical subject differently. The sole exception is the *birth* of the subject which is neutrally depicted no matter the gender and thus henceforth sidelined. The opposite event *death* also becomes slightly more neutral when scaled up in the case of female subjects while staying overly positive – as one would expect from Moravian believes – only in the case of male ones. However, as gbert slightly underestimated the female positivity in the smaller sample, this finding should be taken with a grain of salt. In a similar but opposing manner, gbert overestimated the positivity of female *education or work* although this is less problematic as the distributions per gender converge again when considering the larger sample.

It was mentioned earlier that *family* events may be associated more with female subjects. This is why it is worth mentioning that said domain is depicted neutrally and hence more factive in manually annotated memoirs about boys and men. Although the distributions converge once again when scaled up, the positivity bias seems to affect the latter gender more than the former so that the conclusion may also be true for the larger sample.

With respect to *health* events, which are a very important topic of these sources, it is interesting to note that the interquartile range reaches the neutral range only in memoirs about girls and women. This could once again suggest that said domain is slightly more associated with the female gender. In a similar vein, *housing* changes are also quite important and depicted more positively in the case of this gender. By contrast, the male interquartile range even spans into the negative range – at least in the smaller sample – hence suggesting that more men focused on the negative sides of migration than women did. As this is definitely an interesting finding, it should be further explored as soon as

more men's memoirs become available in digital form. Such a future upscaling would then also enable a closer look at *leisure* events which are rather rare in the sources under investigation here and more likely found in men's memoirs. The distributional disparity could be part of the reason why these events appear to be less positive in the male part of the smaller sample since the more mentions there are, the more likely it is that they differ polarity-wise.

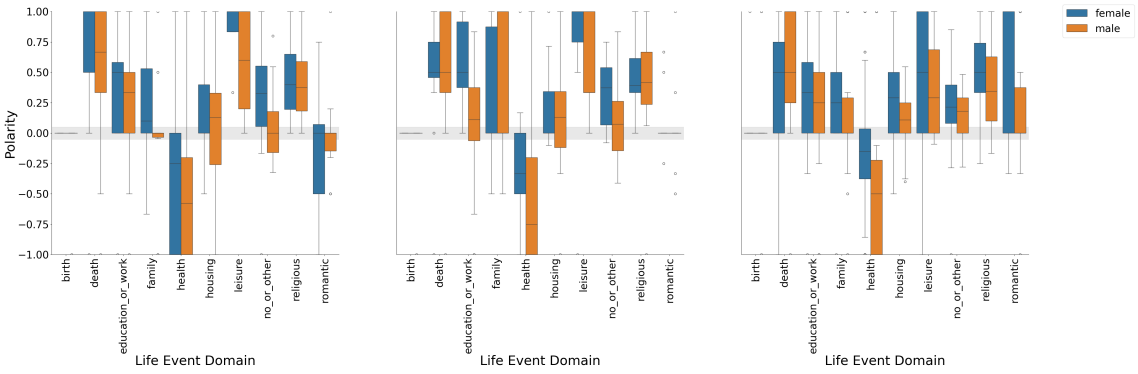

Figure 9.22: Sentiment polarity of life event domains in Moravian memoirs (36 manually annotated left, the same predicted by fine-tuned gbert models middle and all 89 right) grouped by subject gender

Events from the *religious* domain, by contrast, are widespread in the corpus so that the slightly higher positivity found in the case of female Moravians in the manual annotations as well as the fully upscaled sample is worth mentioning. This is even more true when considering that "Moravians believed women were better able to receive, understand, and connect with Christ at a personal and emotional level" in the words of McGuire (2021, 23).

Finally, it was already mentioned in Section 9.2.3 that the few *romantic* events flipped the polarity from negative over neutral to positive when scaled up. However, only the gendered distributions show that the female interquartile range of the smallest sample extents to the positive range although only minimally. In addition, the upscaled version of this gender even extents to maximum positivity whereas the male counterpart does not so that one could conclude romantic events are depicted more positively when experienced by female subjects.

Polarity of life event domains over narrated time

Last not least, a look at sentiment polarity trends per life event domain over the text course sheds even more light on the aforementioned gender-based disparities in the very beginning and at the end of the second third (see Section 9.3.1). To this end, Figure 9.23 shows all trajectories but the one for *birth*. This event was portrayed neutrally and only in the beginning of the narrative so that the trajectory is self-explanatory. Since a chronological rescaling does not add much to the discussion, the respective version is only provided as Figure B.2 in the appendix for reference.

The first disparity observed earlier was a positive female peak in the second bin which coincides with a negative (or neutral) male low. This pattern is observable in basically all nine life event domain trajectories albeit with varying degree. It is most pronounced in the case of the *family* domain and – although postponed by one step – *death*. It is postponed even further in the case of *leisure* activities which are, however, also rather rarely mentioned so that this specific trajectory may be less informative.

Considering the hypothesized low at the end of the second third of the narrative which was less pronounced in memoirs with a female subject, it is worth highlighting that this may also be due to certain event domains. More specifically, the low can actually be observed in descriptions of *health*, *leisure*, and *no or other* events. The latter is especially noteworthy as it accounts for most instances (see also Figure 9.20). Still, this expected female "midlife low" is outweighed by the many positively connoted *housing* changes and *religious* events as well as the less common *death*, *education or work*, *family*, and *romantic* events. Interestingly, with the exception of the *romantic*

domain, the same outweighing affects men's memoirs. However, the lower overall positivity as well as lows in *family*, *leisure*, and to a lower degree also *education/work* events prevent a full outweighing so that the expected low prevails in the case of biographical subjects of said gender.

Figure 9.23: Sentiment polarity of life event domains in 89 Moravian memoirs over narrative time grouped by the gender of the subject as predicted by fine-tuned gbert models

9.3.4 Summary

This second content analysis showed that gender differences manifest themselves in Moravian memoirs in various ways so that some of them are only traceable when combining different analytical scopes. They are also relatable to previous **sentiment analyses** dedicated to English Moravian memoirs where McGuire (2021, 139) observed only slight gender variation for 138 British memoirs from Fulneck while Faull and McGuire (2022, 144) found a higher positivity in nine British women's memoirs from Fetter Lane and twelve American ones from Bethlehem. The same polarity difference was also attested for 89 German memoirs from Bethlehem in the current work where the general trend was observed in trajectories over both narrative and narrated time. This is why *sentiment* seems to be not only a relevant category for Moravian memoirs in general but also one for gender differences in particular. Such differences comprise first and foremost (i) a female high contrasting a male low in the beginning of the narrative and (ii) a slight second low after two thirds of narrative time (see Figure 9.15). The latter one is more pronounced in men's memoirs but becomes less apparent when rescaling to the narrated stages of life.

When relating these findings to **life stage** distributions over the text course, it becomes apparent that women wrote more about (early) childhood whereas men started describing (early) adulthood experiences earlier, and later life stages generally in more detail (see Figure 9.18). Hence, it is likely that the expected childhood high is more pronounced in the case of the former gender due to these different topical foci. In a similar vein, the differences at the end of the second third of narrative time seemingly relate to varying life event domain distributions. More specifically, Moravian women mentioned more positively connoted housing changes and portrayed health issues more neutrally at that point of the narrative on average (see Figure 9.23). As a result, the assumed midlife low is outweighed more than in the case of the opposite gender.

With respect to **life event domains**, it is also worth mentioning that the domain *education or work* was the by far most gender-specific one and is associated with male individuals (see Figure 9.19). *Family*, by contrast, may be a more female-specific domain as related events are depicted more factual and occur more often in non-autobiographical sections in the case of male subjects. Despite notable gender differences, other domains are more difficult to assign. Yet, they are typically more dispersed over the course of male biographies while being grouped slightly more thematically in female ones.

Last not least, it should be highlighted that **non-autobiographical passages** are more common in the case of female subjects so that especially their adult lives are actually described in less detail compared to male ones (see Figure 9.17).

All these findings show that a combination of different classification tasks may not only help to cross-validate findings based on imperfect model predictions but also helps to better approximate covert patterns. Nevertheless, all biographical content analyses presented so far were still only related to one specific century and biographical subgenre. While a cross-temporal investigation stays up to future research, subgenre differences are the focal point of the next content analysis.

9.4 Differences between religious and secular biographies

This final content analysis section compares the major findings of the previous two with another biographical genre – namely entries in the historical German biographical reference work ADB. This second corpus can be expected to build a stark contrast with respect to totality, distance between biographers and subjects, and sociolinguistic backgrounds of the latter while also being similar enough to be analyzed with the fine-tuned gbert models (see Section 4.3). The description of findings and their interpretation will again follow the three tasks with Section 9.4.1 focusing on sentiment polarity, Section 9.4.2 on life stage, and Section 9.4.3 on life event domain patterns. In addition, all subsections but the first relate the respective task-specific findings to the ones described in earlier ones as a means of subsequently adding more perspectives. Finally, Section 9.4.4 concludes with a summary of the most apparent differences between Moravian memoirs and ADB entries before Chapter 10 puts all classification experiments and the content analyses they enabled in a larger context.

9.4.1 Genre differences in sentiment polarity

The overall sentiment polarity distributions per text overlap mostly between both corpora (see Figure 9.24) although ADB entries tend to be more neutral and more diverse polarity-wise. The first difference is in line with the hypothesis that dictionary entries prototypically aim at a neutral factual description unlike conversion narratives. The second may relate to the typical polarity-related attitudes Cockshut (2001, 79) ascribed biographers with respect to their subjects in general. More specifically, the rather positive overall polarity may be a sign of discipleship whereas the negative outliers could be one of hostility. In addition, the predominance of the former is compatible with the claim that most ADB biographers wrote about their peers and sometimes especially about individuals they cherished (Hockerts, 2008, 237–238).

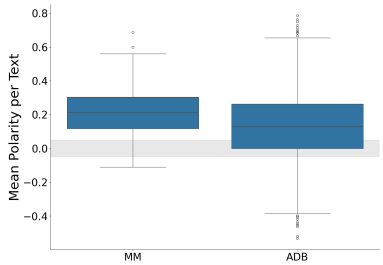

Figure 9.24: Distribution of sentiment polarity in 89 Moravian memoirs (left) and 2630 ADB entries (right) as predicted by gbert*

Polarity trajectories over narrative and narrated time

Even more polarity differences become apparent when considering trends over narrative time (see Figure 9.25). As a matter of fact, the trajectory found in ADB entries is mostly linear with only little variation and thus contrasts the "Rags to riches" pattern with its overall polarity rise found in Moravian memoirs. Hence, only the memoir corpus shows a notable sentiment trajectory which fits the expectation that it may be a genre-specific feature of these sources.

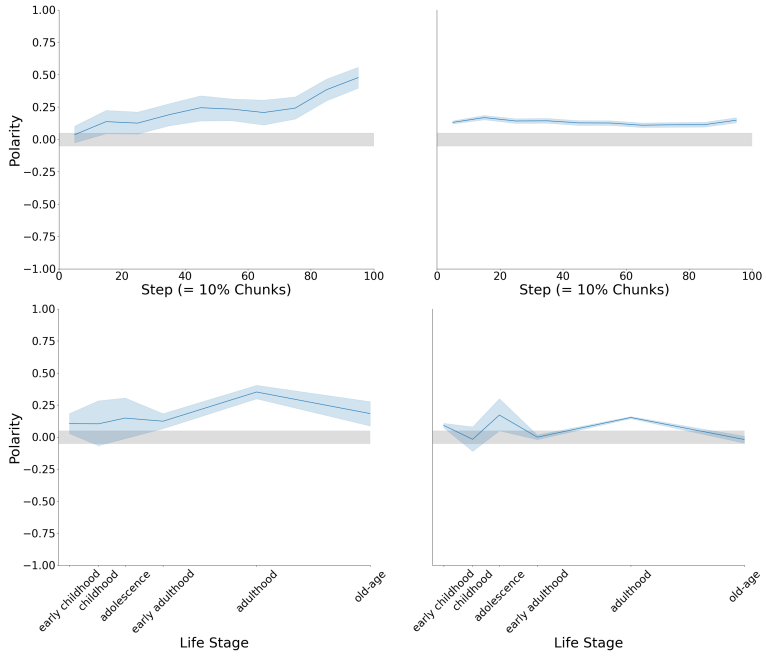

Figure 9.25: Sentiment polarity trajectories in 89 Moravian memoirs (left) and 2630 ADB entries (right) over narrative (top) and narrated time (bottom) as predicted by gbert*

Interestingly, the observed differences diminish after a chronological rescaling to the narrated time where the depiction of (early) childhood and old age appears to be rather neutral while a notable polarity peak occurs only in sections which deal with adulthood in both corpora. Still, the neutrality is actually more pronounced in the ADB, as one would assume. The same is true for the lower adulthood peak.

Only the ADB corpus comprises another notable polarity peak which falls into adolescence and is preceded by a childhood low. Even though this initial polarity fall and subsequent high is missing in the case of Moravian memoirs, it can be related to potential sampling issues. More specifically, the smaller manually annotated Moravian sample was characterized by the opposite pattern but it diminished when upscaling (see Figure 9.7). This implies that the larger Moravian sample comprises at least as many texts with the pattern predominating in ADB entries.

Since the ADB sample is larger and the polarity-wise depiction of adolescence difficult to assess due to the potential sampling issue, it is possible that both corpora are in fact more similar in this regard as it seems. In the case of the Moravian memoirs, this would contradict the hypothesized "fall through adolescence" (McGuire, 2021, 133) but still fit other aforementioned findings. More specifically, (i) adolescence is mostly associated with religious events (see Figure 9.12) and (ii) religious events are positively connoted in these sources. Hence, one could conclude that both hypotheses may be equally valid and just lead to a field of tension with the description of challenges on the one hand, and religious ways to solve them on the other. In this sense, the situation is similar to the one Moravians deal with in the context of illnesses and death.

In summary, this means that both Moravian memoirs and ADB entries depict the life course not only as a whole but also individual life stages in a similar manner polarity-wise. Still, there are noteworthy differences which are mostly a result of the composition or framing so that they can definitely be seen as being related to genre.

9.4.2 Genre differences in life stage distributions

Different to sentiment polarity, the overall distribution of life stages depicted in Figure 9.26 differs considerably between Moravian memoirs and ADB entries. For instance, the latter sources seem to describe *early childhood* in a similar way or even in slightly more detail while basically skipping *childhood* and *adolescence*. The lack of these two stages may relate to the aforementioned bias of the classifier confusing them with (early) adulthood in some cases. However, this difference between the two corpora is also in line with biographical research positing that autobiographical records focus more on earlier life stages than biographical ones (Cockshut, 2001, 78). This is why it is quite likely that the general trend is not (only) an artifact of the classification behavior. Following this interpretation, one could come to the conclusion that the pattern found in Moravian memoirs does in fact not fully falsify the hypothesis of a focus on earlier life stages, it just lacks a point of reference. More specifically, the distributions imply that Moravian memoirs also focus more on later life stages. Yet, they also deal more with earlier ones when compared to other historical sources, such as ADB entries, to which the assumption of Cockshut (2001, 78) that biographers "pass lightly over early years" seemingly fits more. The finding also suggests a lower totality compared to Moravian memoirs which, in turn, contrasts the genre-related assumptions of memoirs being more anecdotal and biographical sources more total (see Figure 4.1).

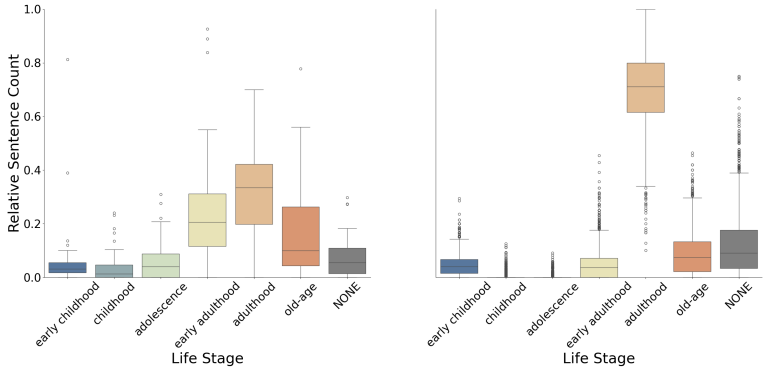

Figure 9.26: Distribution of life stages in 89 Moravian memoirs (left) and 2630 ADB entries (right) as predicted by gbert*

Distributional differences are also observable in the case of *early adulthood* and to a lesser degree *old-age*. While the former difference is difficult to motivate, the latter one may be due to the importance of final illnesses in Moravian memoirs. ADB entries, by contrast, dedicate significantly more sentences to *adulthood* than Moravian memoirs – probably because they want to highlight achievements which are more likely to occur at that stage of life. Similarly, it is possible that the slightly higher amount of *non-biographic remarks* is related to passages that try to put the impact of such achievements in a wider context which fits the academic text genesis.

Distributions over narrative time

When comparing the distributions over the text course (see Figure 9.27), it becomes apparent that both types of life narratives seem to follow the life course chronologically in principle by starting with early childhood and ending with old-age – given the subject did not die earlier. However, as shown above, *adulthood* experiences predominate ADB entries. This is different for Moravian memoirs where they play an increasing role over narrative time, as one would assume in the case of chronological life descriptions.

Other life stages are not only more prominent in Moravian memoirs in general but typically also distributed more over the narrative, such as *early childhood* experiences. These are mostly restricted to the very beginning of ADB entries which implies that they are included due to the totality ambition. Later mentions in the memoirs, by contrast, can be related to the hypothesized thematic priority of autobiographical sources to describe earlier life stages in more detail (Cockshut, 2001, 78).

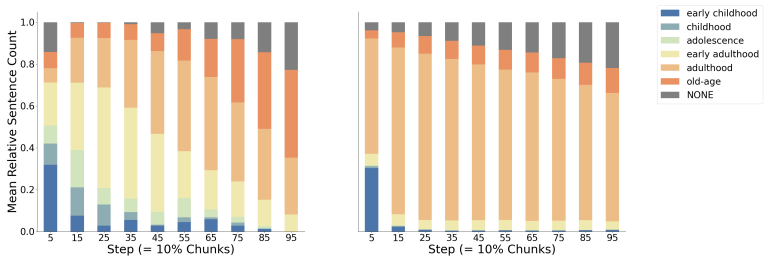

Figure 9.27: Distribution of life stages in 89 Moravian memoirs (left) and 2630 ADB entries (right) over narrative time as predicted by gbert*

Concerning *early adulthood*, it is worth mentioning that related passages appear to be evenly distributed over ADB entries while decreasing over narrative time in Moravian memoirs. The amount of sentences dealing with *old-age*, by contrast, increases in both corpora although less in the case of ADB entries where this last stage of life is described in less detail. In a similar vein, the amount of *non-biographic remarks* increases in both corpora as well albeit that said increase is more linear in the ADB. Hence, they seem to have an overall narrative bracketing purpose in the memoirs under investigation while they are more likely used as a means of offering background information in ADB entries. This fits the more academic nature of the latter corpus.

Finally, it should be added that the overall pattern is characterized by more gradual topical shifts related to life stages in the case of Moravian memoirs which can be related to the "incremental, episodic structure" characterizing memoirs as a genre (Buss, 2001, 595).

9.4.3 Genre differences in life event domain distributions and their polarity

While the overall sentiment polarity distribution was almost identical and the life stage distribution considerably different, Moravian memoirs and ADB entries have different topical foci with regard to life event domains as illustrated by Figure 9.28. More specifically, the merged domain *education/work* appears to be the one major topic of ADB entries. This finding is likely related to the scope of these sources as well as admission criteria which are mostly based on occupational backgrounds (see Section 4.2.1). In a similar vein, it fits expectations that Moravians, by contrast, described *health*, *housing*, and most notably *religious* events in more detail (see Section 9.2.1).

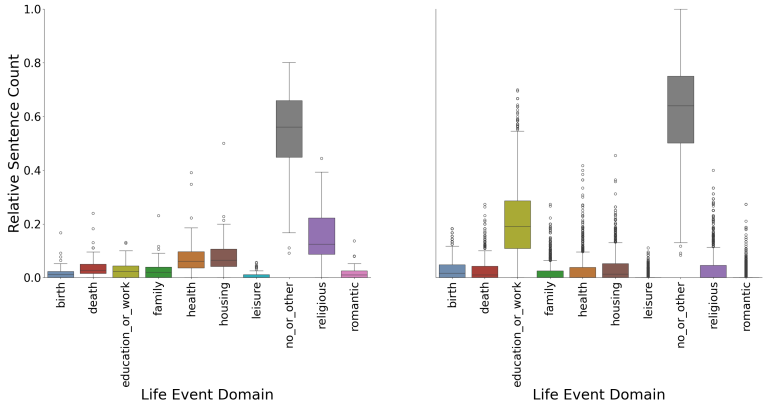

Figure 9.28: Distribution of life event domains in 89 Moravian memoirs (left) and 2630 ADB entries (right) as predicted by gbert*

In contrast to topical foci, both corpora appear to cover domains relatable to basic biographical facts, such as *birth* and *death* of the subject, as well as *family* events in similar detail with ADB biographers writing more about the birth and Moravians slightly more about death and family. With respect to *leisure* and *romantic* events, it is worth highlighting that they occur even less in ADB entries albeit that there are quite a few outliers. This fits the overall pattern that the ADB includes more outliers no matter the life event domain and may be due to certain domains being

niches which become relevant when describing certain lives but are irrelevant in most other cases. Conversely, Moravian memoirs seem to be more homogeneous as somewhat expectable given that they are written as a rite of passage and follow a salvation narrative (see Section 4.1.2).

Distributions over narrative and narrated time

Figure 9.29 shows that the subject's *birth* is only described in the very beginning and only relatable to early childhood in both corpora whereas mentions of her or his *death* increase over narrative and narrated time. This fits the overall chronological life depiction and different ages at death.

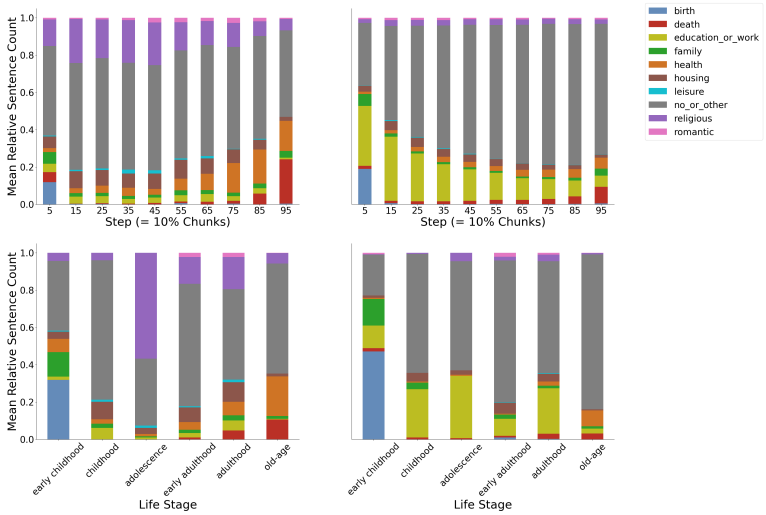

Figure 9.29: Distribution of life event domains in 89 Moravian memoirs (left) and 2630 ADB entries (right) over narrative (top) and narrated time (bottom) as predicted by gbert*

While the merged domain *education or work* is almost equally distributed over narrative time in Moravian memoirs, its share decreases in ADB entries. A likely explanation for this finding could be that the dictionary entries start with a summary of important work in order to highlight the importance of the subject while subsequently adding other topics. This becomes even more likely when considering that events related to education or work are in fact quite equally distributed over descriptions of childhood, adolescence, and adulthood. It also seems plausible that education plays a larger role during childhood and work during adulthood whereas adolescence ranges somewhere inbetween depending on individual differences. However, this is difficult to assess due to the difficulty of distinguishing between these two domains when compiling training data for the classifier (see Section 8.1). The lower proportions in passages dedicated to early childhood or old-age are most probably observable in both sources because fewer people were educated or working during these life stages. An alternative interpretation could be that the decreasing narrative pattern is a result of a framing effect similar to the one found in Moravian memoirs.

One notable turning point appears to be adolescence in both corpora as Moravian memoirs focus on religious events and ADB entries on education or work when dealing with said stage of life. Such a turning point status may also explain the notable drop of the respective domain during descriptions of the subsequent life stage which is early adulthood. A slight focus shift towards *housing* and very few *romantic* events is observable for the latter life stage in both corpora. However, even more topics predominate which are difficult to classify into the ten domains considered here.

With regard to *family* events it is noteworthy that they are most commonly found in early childhood descriptions in both sources and hence also in the beginning of the narrative. They are almost equally distributed over Moravian memoirs, when neglecting gender differences, while being mostly restricted to the first and last bin of ADB entries. The latter distribution is noteworthy as family events can occur at any time in theory but fits the hypothesized first section of said biographical dictionary comprising demographic facts and sometimes even a genealogical abstract (see Section 4.2.1).

Mentions of *health* issues and potential improvements increase over the course of the narrative in both corpora even though they affect different life stages. More specifically, related events appear mostly only when covering old-age (and to a lesser degree adulthood) in the ADB whereas Moravians also mentioned them with respect to almost all life stages but adolescence. This pattern implies that lethal health issues predominate in ADB entries as well – even more so because the lack of mentions during (early) childhood fits the lack of biographical subjects who died at that early stage of life (see Figure 4.4).

As mentioned earlier, *housing* changes and *leisure* activities are more prominent in Moravian memoirs but still vary around their mean proportion in both sources no matter if narrative or narrated time is considered. In addition, events from these domains are less likely to occur in the end of the text or during old-age which may be related to the stronger association of these sections with health issues and death.

Finally, it is notable that *romantic* and *religious* events are more evenly distributed over the ADB narrative than the Moravian one. This implies that they are neither mentioned as often in the former nor placed at a fixed point for narrative text composure purposes.

Polarity of life event domains

Considering the sentiment associated with the different life event domains, it is remarkable that there are actually once again fewer differences between both corpora as shown in Figure 9.30. For instance, the *birth* of the subject is depicted neutrally and hence factual in both sources albeit that the ADB comprises a few positive and negative outliers. This is not the case for the natural counterpart (= the subject's *death*) which is depicted rather positively in Moravian memoirs and quite heterogeneously in ADB entries. The positivity found in the former case can be related to Moravian beliefs as mentioned earlier whereas the ADB distribution is quite intriguing. Since the median falls into the neutral range, the depiction fits a factual description. Nonetheless, the fact that the interquartile range spans from quite negative to rather positive hints at a high degree of variation between entries. Maybe this is related to the hypothesized sign of cherishment (Cockshut, 2001, 237–238) which resulted in an emphasis on the importance (= positivity) in some cases and on the grievous loss (= negativity) in others.

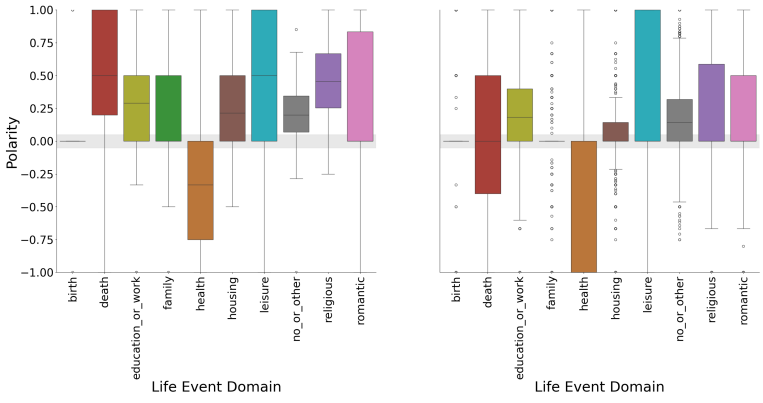

Figure 9.30: Sentiment polarity of life event domains in 89 Moravian memoirs (left) and 2630 ADB entries (right) as predicted by gbert*

The merged domain *education or work* is depicted rather positively although slightly less in the ADB. This is actually quite surprising given the thematic focus on mostly work-related achievements of this biographical dictionary. However, it may have arisen from the field of tension between a factual academic assessment on the one hand and signs of cherishment and discipleship on the other. The more factual reporting in the case of ADB entries may also explain the neutrality with respect to biographical events from the *family* domain.

Conversely, *health*-related events are the only ones which are depicted in mostly negative manner implying that both corpora focus on health issues instead of improvements which is actually

an interesting finding. The same is true for the mostly positive polarity of *housing* changes found in both corpora even though to a lesser degree in the ADB. The positivity of this life event domain is relatable to missionary work in the case of Moravian memoirs but harder to motivate in the case of ADB entries where they could just be reported factually. The similar positive depiction of *leisure* and *romantic* events found in both corpora, by contrast, is likely just related to the fact that said domains are rather positive by nature for the subject who experienced them.

Finally, it is noteworthy to highlight that *religious* events are also depicted rather positively in both sources albeit that the interquartile range extends to the neutral range in the case of ADB entries. The higher polarity found in Moravian memoirs is not so surprising but the positive framing in the ADB definitely is.

When adding the narrative time to the polarity-wise framing, it becomes apparent that, different to Moravian memoirs (see Figure 9.12), the polarity of most domains just meanders around the respective mean in the ADB (see Figure 9.31). One of the few notable exceptions is that there is a positive peak of *death* descriptions at the beginning of the narrative whereas later mentions are more neutral. Conversely, *family* events show a slight negative low in the beginning of the narrative which is outweighed by peaks in the middle and at the end. Still, both domain-specific peculiarities are difficult to assess without a close reading.

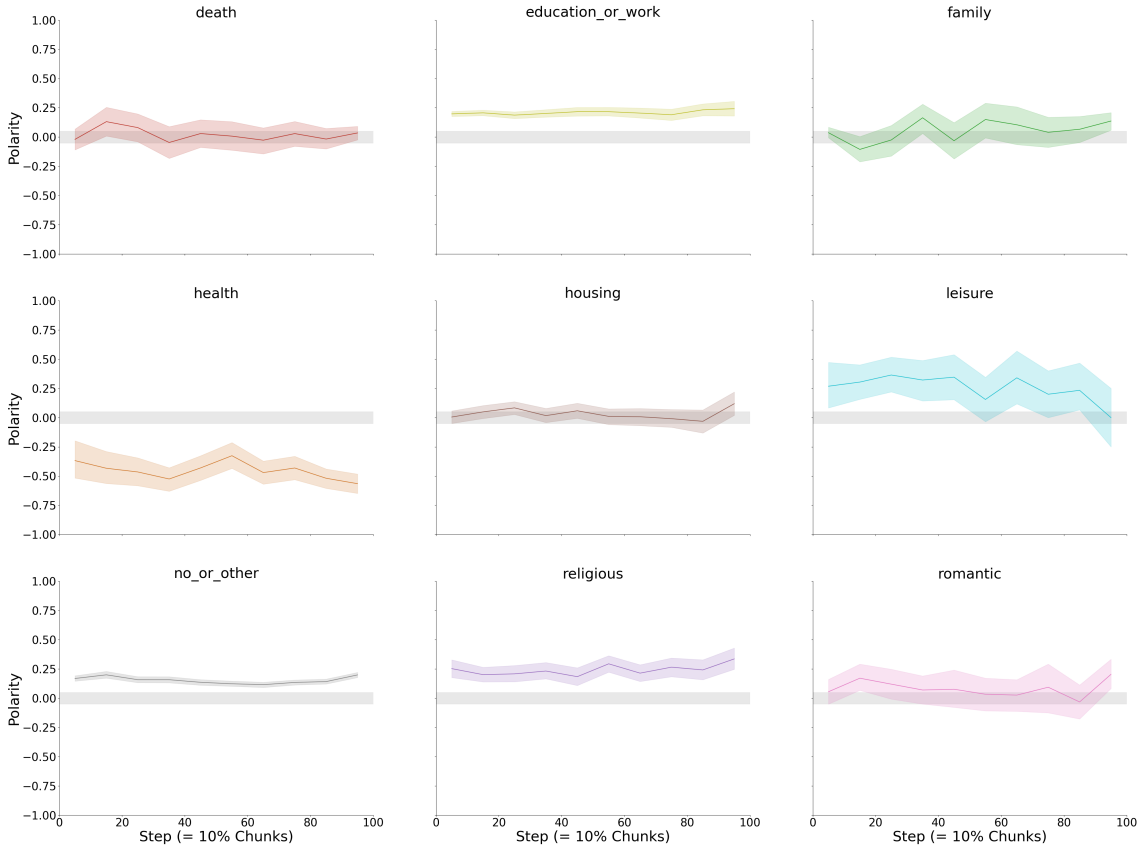

Figure 9.31: Sentiment polarity of life event domains 2630 ADB entries over narrative time as predicted by gbert*

Regarding *housing* changes it is worth mentioning that the observed overall positivity appears to mainly arise from a peak at the end of the narrative on average. Said event domain is therefore actually depicted more factual than previously indicated and hence more in line with expectations. As opposed to this finding, *leisure* events are depicted positively but fall into the neutral range at the end of the narrative. However, they are so rare in this bin that the deviation should not be overestimated. The fact that deviations from the overall patterns are, as a matter of fact, rare in general, supports the conclusion that ADB entries are less composed polarity-wise than Moravian memoirs which is a major (sub)genre-related finding.

9.4.4 Summary

Similar to the content analysis of Moravian memoirs from Bethlehem (see Section 9.2) and potential gender differences in them (see Section 9.3), the comparison with ADB entries yielded many findings which are in line with expectations from more qualitative related work. One such example is the high amount of positive **sentiment polarity** found in both corpora (see Figure 9.24). It likely arises from the salvation narrative in the case of Moravian memoirs and cherished descriptions of achievements in the case of the ADB. However, despite these quite distinct reasons for a rather positive life narrative, the resulting overall sentiment distribution was more similar than one would expect. This is due to the fact that more differences arise when considering polarity trajectories over the text course where ADB entries lack an overall rising pattern (see Figure 9.25). This can be counted as a genre difference between the two corpora because the polarity patterns become more similar again when rescaled to the narrated life course.

Regarding **life stage** distributions, it is worth mentioning that Moravian memoirs are slightly more balanced than ADB entries as the latter focus heavily on adulthood (see Figure 9.26). Thus, it is not surprising that life stages are also less dispersed over the text course in these sources (see Figure 9.27). Furthermore, the scarcity of details about childhood and adolescence in ADB entries makes them less total and more anecdotal than Moravian memoirs which cover all life stages but adulthood in more detail. In addition, the ADB comprises slightly more non-biographic remarks and these seem to serve a different purpose: They become more common over the course of ADB entries – likely to highlight the importance of the described life events – while they are used for more narrative composition purposes in Moravian memoirs.

An analysis of **life event domain** distributions showed that the ADB focuses on education or work whereas Moravian memoirs on religious events and to a lesser degree also health and housing (see Figure 9.28). These topical differences fit the respective scope of these sources. In this regard, it is noteworthy that said difference is especially pronounced in passages about *adolescence*: While Moravians mention challenging experiences as hypothesized, they are outweighed by religious events (see Figure 9.29). In this sense, this stage of life is portrayed as an important turning point of the life narrative even though it is described less wordly than later ones. Similarly, ADB biographers wrote even less about this age group but whenever they did, they highlighted how the respective subject laid the foundations of a later career deemed extraordinary enough for the inclusion in a reference work.

When related to sentiment polarity, it is worth mentioning that religious events are framed more positively in Moravian memoirs than education or work in the ADB (see Figure 9.30). Said latter corpus involves more positive and negative outliers so that the life narration is in fact more heterogeneous polarity-wise. However, both corpora share the remarkable peculiarity that only health-related events are predominantly depicted in a negative way albeit that said domain may also account for positively connoted health improvements. A likely explanation for this disparity is that both corpora rather focus on lethal health issues in the context of this event domain. This is implied by the distribution over both the text and the underlying life course. In this sense, said finding underlines once again that a distant reading interpretation becomes only plausible when considering different interrelated dimensions, such as sentiment, life stages, and life event domains.

Chapter 10

Methodological discussion

Chapters 6, 7 and 8 showed how sentiment polarity, life stage and life event domain classification tasks can be applied to 18th century biographies. Based on these tasks, various content analyses described in Chapter 9 reproduced biographical research results and hinted at patterns not yet reported in related work – at least to the best of my knowledge. In this sense, the main content-related research questions of this work (see Section 1.1) have been answered as follows:

In the case of life stages, certain tendencies are relatable to typical depictions of 18th century lives. Childhood, for instance, was depicted rather positively in memoirs of Moravian women but rather neutrally or even slightly negatively in men's memoirs (see Figure 9.15). This gender disparity can be related to a stronger focus on earlier life stages in texts about women when neglecting sampling particularities, such as the life stage at death. Given that the ADB comprises almost only texts about men, it is worth adding that childhood experiences also appear to be described more briefly and more neutrally in these sources (see Figure 9.25). While especially the latter finding is more difficult to assess due to the imperfect performance of the life stage classifier, it is still likely as it fits expectations from related work.

Life event domains were better identified by the systems used and still there was a similar cross-sample relationship as Moravian men's memoirs and ADB entries seemingly focused on education and/or work whereas women's memoirs dealt more with family-related events (see Figures 9.19 and 9.28). In addition, Moravian memoirs focused more on religious events than ADB entries did and portrayed them in a more positive way (see Figure 9.30). While these findings fit differences in the scope of these sources, the shared feature of depicting only health-related events predominantly in a negative way is a candidate for a more general pattern one can find in descriptions of 18th century lives.

The consistency with findings from related work can be seen as a form of validation of the methodological approach followed here so that some of the methodological research questions have also been already answered implicitly. More explicit answers will follow by reflecting on the methodological implications of the previously presented results. To this end, Section 10.1 deals with the relationship between performance and resource consumption of automatic annotation approaches. Section 10.2 then shifts the focus to explainability approaches by highlighting lessons learned from the application of SHAP and the LRP-based Transformer-Explainability. Finally, Section 10.3 assesses the classifiers trained as part of this work with respect to their applicability to other datasets. As all three sections deal with recommendations for similar research endeavors, they lead over to the concluding remarks in Chapter 11.

10.1 Performances and resource consumption

Expert annotations are generally seen as the most accurate ones albeit that they are not feasible in all project settings. Similarly, larger (classifier) models require more computational resources but actual projects may have an upper limit. This is why it is worth comparing the different automatic annotation approaches of this work both with respect to their performances and their resource consumption. To this end, Figure 10.1 summarizes F1 score distributions per task and system type.

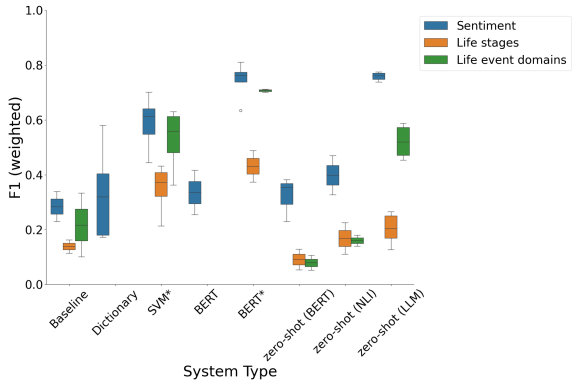

Figure 10.1: F1 scores per system type and task (systems fine-tuned/trained on Moravian data marked with *)

It should be noted that **off-the-shelf dictionaries and BERT models** were only available in the case of the sentiment classification. However, respective systems performed only slightly better than the random and majority baseline on average with some dictionaries performing even worse. This finding is noteworthy because many sentiment analyses in the field of digital humanities are based on such systems which raises validity questions (see Section 3.1.4).

Similarly, many **zero-shot approaches** yielded comparably low performances no matter the task at hand even though a parameter size effect was observable. More specifically, BERT models performed worst in the zero-shot setting while NLI models were similar to the baselines, and the two LLMs performed best in most of their runs. This size effect was the lowest in the case of the life stage categorization where the evaluated LLMs underperformed compared to the baselines in a few runs and it was actually the smaller model which performed better (see Table 7.3).

A positive outlier worth mentioning was the sentiment classification task where most LLM runs were among the most performant settings. This task-specific disparity is in line with the observation of Bamman et al. (2024, 495) that "prompt-based LLMs [are] competitive with traditional supervised models for established [...] tasks, but lag behind supervised models for de novo tasks".

Regarding "traditional supervised models", it is worth highlighting that **SVMs** fitted on distributional features of data from the target domain performed second-best in most settings. Nevertheless, there is a remarkable F1 score range when only accounting for different vectorizers, n-gram ranges, and analyzer levels (see Section 6.2.2). Thus, the performed grid search which aimed at a parameter optimization was worth the effort and is advisable to similar projects.

It should be underlined that the evaluated **fine-tuned BERT models** in general and the basic German one in particular outperformed most other evaluated systems by a margin. The performance gain compared to the best LLM setting was the lowest for a sentiment and also quite low compared to the best SVM setting for a life stage classification. Since the fine-tuning used default parameters as suggested in related work, further performance gains are in the realms of possibility.

In general, it is noteworthy that the overall F1 score decline between the sentiment polarity, life event domain and life stage classification tasks do not fit the amount of classes (= three, ten, and seven) involved. Instead, it appears to hint at the fact that the three sentiment categories are more coherent on the base of textual features than (some of) the life event domains. This is, for instance, related to the fact that it makes a difference whether the event experiencer is the biographical subject or not (see Section 3.3.1). In a similar vein, the life stage categories can be seen as the least coherent ones – not only because of ambiguous cases but also because it is disputable whether one should subdivide the life course this way (see Section 3.2.3).

As most life stage label confusions were in fact only off-by-one (see Section 7.3.2), it can be counted as a lesson learned that classification tasks which are based on the projection of a rather continuous scale like *age* to discrete categories add less coherent edge cases. However, there are also ways to account for this issue, such as a more in-depth error analysis. In addition, it should not be forgotten that it depends on the downstream task at hand how much of an issue off-by-one errors are. As an example, Vauth et al. (2021, 340) also observed many confusions of this kind in the case

of their narrative event classification but argued that they somewhat average out when aggregating to trajectories over the text course. Hence, although one should keep these issues in mind, it is likely that they also had less impact on the life stage distributions analyzed in the previous chapter than one would assume based on raw F1 scores.

Resource consumption

The evaluated systems differ with respect to computational resource requirements. However, it is difficult to assess the exact requirements which is why they are approximated by (i) the run time and (ii) the respective hardware (= CPU or GPU) in the following. The results are grouped by system type and application scenario in Table 10.1 (and shown for each individual system in Table C.1).

System Type	Training	Inference		
		Test	MM	ADB
Dictionary	–	00:00	00:00	–
SVM*	00:04	00:01	00:03	02:40
BERT	–	00:40	03:30	–
BERT*	*09:12*	00:53	03:50	02:55:27
Zero-shot (BERT)	–	06:11	–	–
Zero-shot (NLI)	–	24:13	–	–
Zero-shot (LLM)	–	*02:36*	*10:19*	–

Table 10.1: Mean runtimes per system type for training with 1768, testing with 442, and inference on 2210 Moravian and 115,486 ADB sentences (in hours, minutes and seconds in the form (hh:)mm:ss; systems fine-tuned/trained on Moravian data marked with *; GPU usage in italics)

As expectable, **dictionary lookups** required the least amount of resources but also yielded the worst results. **SVMs**, by contrast, were much more performant while still only requiring a few seconds on a standard CPU for training and inference on Moravian data. In addition, an upscaling to the 115 thousand ADB sentences took less than three minutes so that this approach is feasible for projects with limited resources.

BERT models required more resources but still only a standard CPU for inference. More specifically, labeling the 442 test instances took a little less than a minute but an upscaling to the ADB sample almost three hours. However, off-the-shelf models yielded low performances so that respective models appear to be only applicable to specialized (historical) domains when fine-tuned. Unlike inference, which was possible on a standard CPU, the fine-tuning itself required GPU access but took only a little more than nine minutes per model on a T4 cloud GPU (= three per epoch). Hence, the high performance gain has outweighed the increase in resource consumption.

The difference between training and inference requirements is important because the former is usually conducted less often than the latter. Similarly, the increase in resource consumption during training can be seen as less problematic in the case of underresourced textual domains, such as historical biographical sources, since a training with more samples requires even more computational resources. Therefore, the question of how many training instances are actually necessary for satisfying performances is relevant for domains with larger digital support. This is even more true given that a smaller training set does not necessarily lead to performance losses (Mosbach, 2023; Brookshire, 2026) so that it is up to individual projects to balance between performance and resource consumption. In this work, said field of tension was resolved by neither exploring parameter nor architectural optimizations of transformer models.

With respect to **zero-shot approaches**, there is a clear relationship between parameter count and resource consumption with LLMs even requiring GPU access for inference unlike the smaller BERT and NLI models. Furthermore, the runtime of zero-shot models was influenced by the amount of classes, as shown in Table 10.2. More specifically, distinguishing between three sentiment labels consistently required the least amount of resources while the seven life stages required more, and

the ten event domains the most. Regarding NLI models, it is in line with the observation of Laurer et al. (2024, 95) that an inference takes considerably longer than with fine-tuned BERT models while yielding considerably worse results. Hence, their claim that this zero-shot approach is not an advisable one in cases where at least 2000 training samples are available or compilable is supported. Similarly, the vast amount of resources required by LLMs contrasts the rather low performance in tasks which are not as widespread as a sentiment classification so that their usage seems to also not be advisable for similar research endeavors.

Zero-shot system	**Type**	**Test Inference**		
		Sentiment	**Life Stages**	**Life Event Domains**
bert_110M	BERT	03:53	–	–
bert_multilingual_110M	BERT	03:08	07:32	08:20
gbert_110M	BERT	03:00	07:15	10:07
mDeBERTa-nli_279M	NLI	05:30	17:16	20:00
bart-nli_406M	NLI	16:25	37:00	49:04
Gemma3_27B	LLM	*02:15*	*02:21*	*02:23*
Nemotron_Ultra_253B	LLM	*02:34*	*02:56*	*03:04*

Table 10.2: Mean runtimes of zero-shot approaches per task when applied to 422 test samples (in minutes and seconds in the form mm:ss; parameter count prefixed with _; GPU usage in italics)

In summary, a supervised upscaling from a small manually annotated dataset to a larger one from a quite similar textual domain and time frame seems to be a very promising approach. System-wise, "traditional" machine learning approaches like SVMs may be a good starting point whereas a transformer fine-tuning should be preferred whenever GPU access is available for training.[25]

10.2 Explainability of transformer-based classifiers

With respect to the methodological research questions of this work (see Section 1.1), the previous section showed once again that especially fine-tuned transformer models are able to (i) overcome the domain-sensitivity of sentiment polarity, (ii) infer some of the life stages from biographical descriptions – albeit with notable issues when trying to discriminate between subsequent stages –, and (iii) group described life events by the domain they concern. Especially in the first case, transformer-based classifiers outperformed more transparent rule-based systems so that the transparency issues raised in related work come into play. More specifically, Kim and Klinger (2021, 11) assumed that worse performing rule-based systems are actually preferred in projects from the fields of digital humanities for said reason and Van Atteveldt et al. (2021, 134) observed similar tendencies in the computational social sciences.[26]

In order to mitigate transparency issues, the two explainability approaches *SHAP* (Lundberg and Lee, 2017) and *Transformer-Explainability* (Chefer et al., 2021) were applied in this work. As shown in Section 6.3.3, both are able to explain which subword tokens in a given instance affected label attributions the most and hence answer the research question of whether it is possible to explain the classification behavior despite a "black box" nature.

However, it should also be noted that it is difficult to compare the "performance" of explainability models due to their "black box" nature. This is why Adadi and Berrada (2018, 52154) stated that "formalized rigorous evaluation metrics" are a notable research gap in the field of explainable artificial intelligence. Nevertheless, both approaches are established ones and hence deemed applicable – even more so as they yielded quite similar results in the case of the sentiment classification (see Section 6.3.3).

[25] As a labeled dataset is necessary for the fine-tuning, it can also be used to assess whether off-the-shelf (BERT) models are already applicable so that the hardware requirement may become obsolete in other work.

[26] It is unclear to which extent these observations also apply to LLMs given their current ubiquity.

Still, there are differences between the two explainability approaches. SHAP, for instance, assigns negative values to subword tokens contradicting the given label while Transformer-Explainability does not so that only the former can be used to illustrate textual features of this kind. However, it also requires considerably more computational resources as shown by the fact that the four examples in Figure 6.7 where analyzed in two minutes whereas Transformer-Explainability needed only six seconds. This drawback of SHAP appears to be especially pronounced when applied to BERT models (Zielinski et al., 2023, 145) which is why Transformer-Explainability was preferred in most sections of this work.

In the case of the life stage classification, the conducted explainability analyses helped to motivate some of the decision-making albeit that many ambiguous cases arose when aiming at a global explainability on the base of the test dataset (see Section 7.3.3). This could either be related to the difficulty of the task or hint at a low performance of the explainability approach. The former appears to be more likely given that the respective analysis of the life event domain classifier were more in line with expectations (see Section 8.3.3). Still, it remains difficult to completely rule out confirmation bias. Moreover, it should be noted that even if a global explainability analysis of the test dataset yields plausible most predictive features, it is impossible to predict the classification behavior in cases that lack them. Hence, it can also never be ruled out that some label predictions on the one hand and some relevance scores on the other are just wrong.

Yet, these (potential) limitations should neither be overestimated nor be seen as outweighing the gains of adding explainability to deep learning classifiers: They are, for instance, especially useful in the case of "unexpected decisions" according to Adadi and Berrada (2018, 52142). This specific use case played a role in Section 9.2.3 where the relevance scores suggested that gbert has learned historical as well as regiolectal word forms in the case of various variants of the verb *heiraten* 'to marry'. While such a finding can be related to the research question of whether the model overcame scarcity issues of historical sources, it first and foremost serves the purpose of increasing transparency and by extension also trustworthiness. This is due to the fact that it is easier to trust results gained from a model which labels sentences not only consistently but also on the base of textual features a human annotator would also consider relevant.

In a similar vein, explainability methods may hint at biases in the training dataset (Brookshire and Reiter, 2024, 95–96) so that their identification may serve debugging purposes which in turn may lead to further model improvements (Adadi and Berrada, 2018, 52143). One such model improvement possibility is to deliberately oversample sentences without the previously identified biases in a second fine-tuning step. However, this approach comes at the cost of additional computational resources which is why it is also worth considering less resource-intensive options that aim at an increase in performance with respect to downstream analyses. To this end, a Python pipeline was developed within the context of this work which enables human-in-the-loop workflows where notable outliers in automatic annotations are first identified through distant reading and then manually assessed (and corrected if necessary) in a close reading step supported by an explainability model (Brookshire, 2025).

In spite of all these possible applications, it should, however, also be noted that any explainability analysis is only able to approximate a more transparent decision-making which can (up until now) only be gained from more traditional models based on rules.

10.3 Applicability to other datasets

The **sentiment** classification task yielded the highest scores when only considering supervised learning and zero-shot settings. The fine-tuned BERT models and to a lesser degree also the SVMs based on distributional features are therefore the most apt ones to be applied to other datasets. Still, it was already mentioned by Brookshire and Reiter (2024, 98) that "we expect our fine-tuned models to perform worse when applied to strongly deviating domains" which is also true for the SVMs. However, this is a somewhat expectable limitation since off-the-shelf sentiment models have mostly been trained on contemporary social media and customer review data which is why they tend to perform worse on humanities datasets (Klähn et al., 2025). This expected behavior was

also observed here so that a cross-domain applicability limitation appears to be a typical feature of context-sensitive tasks, such as a sentiment classification (see Section 3.1.4).

It is worth underlining that sentiment models in general are not only characterized by biases towards predicting a certain label but also by their potential ethical risk of basing their decision-making on factors like age, gender, and racial stereotypes (Venkit et al., 2023, 13749). As a debiasing measure, personal names have been masked in the case of training and test samples here (see Section 5) following the respective suggestion by Hube et al. (2020, 259). However, this practice does not rule out that the model(s) may have learned to associate other references to human beings with sentiment polarity. These potentially harmful stereotypes may, for instance, relate "to Native Americans or immigrants with non-European origins" and "should be seen in the given historical context" as already noted by Brookshire and Reiter (2024, 99). Hence, predictions of the fine-tuned transformers and fitted SVMs should be analyzed accordingly when applied to datasets that contain respective phrases and should not be taken as a means of fostering derogatory language. To give an example, the model(s) may be apt for quantitative analyses of verbal devaluation found in Moravian texts like the ones Lasch (2024) conducted.

Regarding the **life stage** classification, it should be noted that Giele and Elder (1998, 9–10) identified inter-cultural and inter-temporal differences as being very influential factors for individual differences in life course trajectories. Thus, the applicability to other sources may be even more restricted for the model(s) trained on this task. Still, this task enabled a chronological rescaling from the narrative to the narrated time in the content analysis chapter where it led to noteworthy additional findings. This is why the actual model(s) may not be as easily applicable to other datasets but the underlying methodological approach definitely is.

Section 8.3.3 showed that gbert learned Moravian-specific vocabulary when fine-tuned for a **life event domain** classification. One notable example is the strong association of the tokens *Chor(häuser)* 'choir (houses)' with educational or work-related events. This makes sense as the choir system was a very important aspect of 18th century Moravian communities (see Section 4.1.1). In addition, the transformer model has seemingly also learned (historical) graphematic and regiolectal word forms like in the case of the verb *heiraten* 'to marry' (see Section 9.2.3). Therefore, said model is also highly adapted to this specific textual domain which in turn decreases its applicability to sources from other domains and time frames. Nevertheless, this high domain-specifity makes the model(s) apt for further upscalings to the many more Moravian sources that have already been digitized in the form of fulltext transcriptions or can be expected to be in the near future. One example corpus for the former are 18th century Moravian travel diaries (Prell, 2022) and one for the latter is the still ongoing project Moravian lives (see Section 4.1.3).

Chapter 11

Conclusions

This work showed how life events and their polarities can be traced in 18th century biographical sources with transformer-based language models. In addition, it showed how both concepts can be related to age-based subdivisions of the life course commonly called *life stages* (see Section 3.2.1). This was achieved by manually annotating 2210 sentences from 36 Moravian memoirs which have been transcribed digitally in related work (see Chapter 5). 1768 sentences from said corpus were used for training and 442 for testing in three classification experiments. These experiments underlined that such a comparably scarce dataset is an apt one for transformer models to learn task and domain-specific features in a fine-tuning setting so that they outperform most off-the-shelf approaches by a margin. Especially in the case of the ternary **sentiment classification**, the result was with an F1 score of .81 for the best system (= the basic German BERT model gbert) in the state-of-the-art range (see Section 6.3.1).

Similarly, the same model yielded an F1 score of .71 when fine-tuned for a **life event domain classification** (see Section 8.3.1) which is comparable to the mean F1 score of .72 from a survey of event type classifiers (Xiang and Wang, 2019, 173130). Still, both scores are slightly less comparable since the survey considered different event type classes (= the ACE standard) which have been shown to not quite match typical requirements from historical studies as problematized by Sprugnoli and Tonelli (2017).

An even more pronounced comparability issue arises in the case of the **life stage classification** as this task is a novel one. Still, the rather low F1 score of .49 for gbert fine-tuned to this specific task was still somewhat expectable. This is due to the fact that there are many generic phrases which are not as easily relatable to one life stage or another (see Section 7.3.2). In this sense, the overall F1 score decline between the three tasks fits the respective decline in coherence of the categories involved (see Section 10.1).

Most misclassifications of the life stage classifier were in fact only "off-by-one". They can hence be motivated with individual differences in human development even though an external **explainability model** failed to make the more general decision-making process and potential biases more understandable (see Section 7.3.3). Conversely, in the case of the other two fine-tuned gbert models, explainability analyses underlined that gbert indeed learned to identify subword tokens a human would also relate to the respective task at hand. In addition, it seemingly learned to follow more implicit annotation guidelines for resolving ambiguous cases (see Section 6.3.3). In a few cases, the majority class of the training dataset was preferred implying that this class may be seen as some kind of default value. However, training data support can not explain the significantly higher performance compared to other approaches in all three tasks. This is even more true given that the fine-tuned transformers also showed strong performances for a few scarcely supported classes, such as birth descriptions with only 21 training instances (see Section 8.3.2).

In summary, this means that transformer models in general and gbert in particular are apt for an upscaling from a small(er) manually annotated corpus to a larger one which answers most of the methodological research questions of this thesis. Still, it should also be mentioned that this is not the case for off-the-shelf sentiment models fine-tuned with modern-day texts. This disparity is in line with the observation of Klähn et al. (2025) that these models struggle with humanities datasets.

However, as the amount of training samples was a smaller issue here, a similar upscaling approach may also resolve the issue in other future digital humanities projects with a similar scope.

Content-wise, Chapter 9 illustrated that the 89 analyzed Moravian memoirs from 18th century Bethlehem, Pennsylvania, follow notable patterns which also encompass gender differences. Some of the latter were more covert which may explain the mixed results in related work (Mettele, 2007; Faull and McGuire, 2022). In the case of women's memoirs, the most pronounced ones were (i) a higher overall sentiment polarity and (ii) more non-(auto)biographical passages. They also focused more on earlier life stages, and their male counterparts on later ones. Due to further interrelations with different topical foci, this has implications for the sentiment trajectory over the narrative as cross-validated with a TF-IDF analysis in Section 9.3.3. More specifically, the beginning of women's memoirs was mostly characterized by positively connoted religious events while men described work (and housing) events in a rather negative manner. These findings underline the added value of combining multiple perspectives as approximated with three classification tasks since such a deep distant reading understanding would not have been achieved with one task alone. This supports the methodological critique that sentiment trajectories over narrative time are neither the only way to approximate "plot" patterns (Gius and Vauth, 2022, 2) nor the best one when considering narratological theory (Rebora et al., 2023, 7).

Still, it should be highlighted that the polarity-wise depiction of life stages was not only quite similar between women's and men's memoirs (when neglecting opposing trends during childhood/adolescence) but also between Moravian memoirs and entries in a 19th century biographical dictionary. This is noteworthy because other distributions were quite distinct (see Section 9.4). For instance, the polarity trajectory over the narrative was mostly linear in the latter case suggesting that only Moravian memoirs portrayed lives as salvation narratives. Hence, this disparity between narrative and narrated time implies that said difference is related to the text composition.

In a similar vein, there were notable topical differences with the dictionary entries focusing almost exclusively on adulthood experiences on the one hand and on education or work events on the other. Interestingly, these patterns are more similar to men's memoirs and therefore fit the gender distribution in the biographical dictionary under consideration.

By contrast, the high proportion of non-biographical remarks found in the latter sources increases the similarity with women's memoirs even though said passages seemingly serve different purposes: In the dictionary, they appear to be used in order to put achievements of the subject in a broader context whereas Moravians rather used them as a means of balancing out negative sides of the final phases of life. In this sense, such remarks can be claimed to foster genre differences so that their closer analysis may be one starting point for subsequent research endeavors. How future work can build up on this one will be elaborated further in the following.

11.1 Main contributions

Besides the identification of the aforementioned notable patterns in two historical biographical corpora, the main contributions of this thesis are (i) labeled datasets, (ii) trained classifiers, and (iii) a generic explainability pipeline:

The manually **annotated Moravian datasets** can be reused for further close and distant reading approaches in studies dedicated to Moravian sources. As two annotation layers (= life stages and life event domains) concern novel tasks, they may be especially apt for analyses, such as the depiction of childhood on the one hand or health events on the other. The first one is not only worth investigating due to the observed gender differences but also because the 18th century itself is characterized by a "discovery" of said life stage (Taylor, 2016). In the case of health events, it was remarkable that they were mostly portrayed in a rather negative manner so that the focus appears to rather lie on health *issues* than on health *improvements*. This applied to the biographical dictionary as well so that it may be a more general pattern found in historical biographical sources.

In order to enable future work that identifies even more such patterns, this work contributed **fine-tuned BERT models and fitted SVMs**. However, it should again be underlined that they can be expected to be most apt for textual sources as similar to the ones analyzed here as possible. This

is due to the high domain-sensitivity of the approaches (see Section 10.3). One way for assessing the applicability a priori is a PCA of the underlying vectorization/embedding (see Section 4.3) albeit that the compilation of a (small) test dataset may be even more advisable.

Finally, a **generic Python module** for any combination of a text classifier and explainability model was developed by Brookshire (2025). This "side product" of the current work is able to mitigate transparency issues of the fine-tuned transformer-based language models which performed best here. It only requires annotated data to be in a tabular format with the column headers `source`, `n`, `text` and `label`.[27] It then adds one column per class (probabilities) automatically and highlights the importance of individual (subword) tokens (see e.g. Figure 9.11). Since the underlying color palette is fully customizable although intensity differences are reserved for illustrating feature relevance (see Figure 6.7), it can easily be adapted to different tasks. As shown in Section 9.2.3, one use case is a deeper understanding of the decision-making in peculiar cases. Another one could be the identification of unwanted biases in training samples that allow for some kind of "informed augmentation" when compiling a less biased future fine-tuning dataset.

11.2 Future directions

Since this work dealt with multiple tasks combined in downstream analyses, training a multi-task model might be a reasonable next step in computational linguistic research projects. A classification experiment of this kind could test whether a combination leads to better results in order to verify the presumed interrelations between the tasks mentioned earlier. This was not yet done since only little related work suggested correlations between sentiment, life stages, and life event domains so that one focus of this work was to identify them first. It is, however, possible that the correlations are even more domain-specific as they were more pronounced in the Moravian memoirs under consideration than in the case of the dictionary. Thus, a combined model might be less applicable to similar datasets than the individual models even though this remains to be proven in future research which could profit from the choice of annotating the exact same sentences for all three tasks.

With respect to the content analyses conducted within this work, future work could strengthen the validity as soon as more Moravian memoirs in particular and more historical biographies in general become available in the form of fulltext transcriptions. Analyses of such larger samples could, for instance, lead to a formal typology of patterns found in Moravian memoirs which tries to resolve the field of tension between individuality and norm described by Faull (2021, 217–218).

In a similar vein, an increase in metadata readily available or at least inferable would allow for the analysis of more demographic variables other than *gender* with the systems developed within this thesis. One such variable worth investigating is *author age* because psycholinguistic research has suggested certain correlations with sentiment polarity. More specifically, older individuals have been shown to "use fewer negative emotion words and more positive" ones (Pennebaker and Stone, 2003, 299). This is also noteworthy as it matches implicit biases of human annotators in crowd-sourcing projects like the one of Flekova et al. (2016, 851). Due to the unclear authorship of Moravian memoirs, the actual author age may, however, be difficult to assess. The age of the deceased one on behalf of whom the memoir was composed could be a possible coarse-grained approximation. Some other demographic variables worth analyzing on the base of larger samples are *occupational groups* or more generally *social classes* on the one hand and *regional differences* on the other. Finally, a look at *diachronic differences* could be another interesting goal of future research albeit that such work will face the difficulty of deciding how to ensure comparability given the domain-specific nature of the tasks and the automatic annotation approaches.

In the end, both this and similar future work make even more historical life events and their polarities traceable so that more historical perspectives become uncovered as biographical sources play a larger role in (digital) historical research.

[27] The first two columns are usually not used in text classification datasets and also removed from the figures of this thesis as an anonymization effort. Still, these columns enable a backtracing of the text instance to its position in the original corpus which is useful for most application scenarios.

Bibliography

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Shivaji Alaparthi and Manit Mishra. 2021. BERT: a sentiment analysis odyssey. *Journal of Marketing Analytics*, 9:118–126.

Ruth Albrecht. 2022. Biographik in der Frühen Neuzeit. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 341–344. J.B. Metzler, Stuttgart, Germany.

Hafiz Muhammad Usman Ali, Qaisar Farooq, Azhar Imran, and Khalil El Hindi. 2025. A systematic literature review on sentiment analysis techniques, challenges, and future trends. *Knowledge and Information Systems*, 67(5):3967–4034.

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable Annotation of Events and Event Relations across Domains. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 10–20, Santa Fe, NM, USA. Association for Computational Linguistics.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Craig D. Atwood. 2004. *Community of the Cross: Moravian Piety in Colonial Bethlehem*. Pennsylvania State University Press, University Park, PA, USA.

Matthias Aumüller. 2022. Narrativität. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 23–27. J.B. Metzler, Stuttgart, Germany.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. In *Proceedings of the Computational Humanities Research Conference 2024*, pages 494–527, Aarhus, Denmark. CEUR Workshop Proceedings.

David Bamman and Noah A. Smith. 2014. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Barbara Becker-Cantarino. 2016. Rechenschaft und Kontrolle: Herrnhuter Lebensläufe aus der ‚Indianermission' in Nordamerika ca. 1760–1800. *Aufsatzpraktiken im 18. Jahrhundert*, 28:173–190.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. 2025. Llama-nemotron: Efficient reasoning models.

Laura Bernardi, Johannes Huinink, and Richard A. Settersten. 2019. The life course cube: A tool for studying lives. *Advances in Life Course Research*, 41:100258.

Ágoston Zénó Bernád and Maximilian Kaiser. 2018. The Biographical Formula: Types and Dimensions of Biographical Networks. In *Proceedings of the Second Conference on Biographical Data in a Digital World*, pages 45–52, Linz, Austria. CEUR Workshop Proceedings.

André Blessing, Andrea Glaser, and Jonas Kuhn. 2015. Biographical data exploration as a test-bed for a multi-view, multi-method approach in the Digital Humanities. In *Proceedings of the First Conference on Biographical Data in a Digital World*, pages 53–60, Amsterdam, the Netherlands. CEUR Workshop Proceedings.

Orville G. Brim. 1980. Types of Life Events. *Journal of Social Issues*, 36(1):148–157.

Patrick D. Brookshire. 2025. Warum wird was wie klassifiziert? Scalable Reading + Explainable AI am Beispiel historischer Lebensverläufe. In *Book of Abstracts - DHd 2025*, pages 535–537, Bielefeld, Germany.

Patrick D. Brookshire. 2026. Finding and Classifying Names – Challenges and Opportunities for Fine-tuned Transformers. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*, 12:7–26.

Patrick D. Brookshire and Nils Reiter. 2024. Modeling Moravian Memoirs: Ternary Sentiment Analysis in a Low Resource Setting. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 91–100, St. Julians, Malta. Association for Computational Linguistics.

Steven Brown and Carmen Tu. 2020. The shapes of stories: A "resonator" model of plot structure. *Frontiers of Narrative Studies*, 6(2):259–288.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Helen M. Buss. 2001. Memoirs. In Margareta Jolly, editor, *Encyclopedia of Life Writing. Autobiographical and Biographical Forms*, volume 2, pages 595–596. Fitzroy Dearborn Publishers, London, UK.

Stephanie Böß. 2016. *Gottesacker-Geschichten als Gedächtnis. Eine Ethnographie zur Herrnhuter Erinnerungskultur am Beispiel von Neudietendorfer Lebensläufen*. Number 6 in Studien zur Volkskunde in Thüringen. Waxmann, Münster, Germany.

Tommaso Caselli and Roser Morante. 2018. Systems' Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Tommaso Caselli and Ahmet Üstün. 2019. There and Back Again: Cross-Lingual Transfer Learning for Event Detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, pages 77–84, Bari, Italy. CEUR Workshop Proceedings.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, Nashville, TN, USA. IEEE.

Yanqing Chen and Steven Skiena. 2014. Building Sentiment Lexicons for All Major Languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, MD, USA. Association for Computational Linguistics.

John A. Clausen. 1998. Life Reviews and Life Stories. In Glen H. Jr. Giele, Janet Z. and Glen H. Jr. Elder, editors, *Methods of Life Course Research. Qualitative and Quantitative Approaches*, pages 189–212. SAGE Publications, Inc., Thousand Oaks, CA, USA.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13, Lisbon, Portugal. University of Zurich.

A. O. J. Cockshut. 2001. Autobiography and Biography: Their Relationship. In Margareta Jolly, editor, *Encyclopedia of Life Writing. Autobiographical and Biographical Forms*, volume 1, pages 78–79. Fitzroy Dearborn Publishers, London, UK.

Katherine Coles and Julie Gonnering Lein. 2013. Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research. In *Digital Humanities 2013. Conference Abstracts*, pages 150–153, Lincoln, NE, USA.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Alessandra Corvonato. 2021. Re-Evaluating GermEval 2017: Document-Level and Aspect-Based Sentiment Analysis Using Pre-Trained Language Models. Master's thesis, LMU, München, Germany.

Sanjiv R. Das and Mike Y. Chen. 2007. Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9):1375–1388.

Laleh Davoodi and József Mezei. 2022. A Comparative Study of Machine Learning Models for Sentiment Analysis: Customer Reviews of E-Commerce Platforms. In *35th Bled eConference Digital Restructuring and Human (Re)action, Conference Proccedings*, pages 217–231, Bled, Slovenia. University of Maribor Press.

Ashley Dennis-Henderson, Matthew Roughan, Lewis Mitchell, and Jonathan Tuke. 2020. Life still goes on: Analysing Australian WW1 Diaries through Distant Reading. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 90–104, Online. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 837–840, Lisbon, Portugal. European Language Resources Association.

Detlev Dormeyer. 2022. Religionswissenschaft. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 519–525. J.B. Metzler, Stuttgart, Germany.

Keli Du and Katja Mellmann. 2019. Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung. Ein Pilotprojekt. *DARIAH-DE Working Papers*, 32.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Katherine Elkins. 2022. *The Shapes of Stories. Sentiment Analysis for Narrative*. Elements in Digital Literary Studies. Cambridge University Press, Cambridge, UK.

Frederik Elwert and Simone Gerhards. 2017. Tracing Concepts – Semantic Network Analysis as a Heuristic Device for Classification. In Tanja Pommerening and Walter Bisang, editors, *Classification from Antiquity to Modern Times*, pages 311–338. De Gruyter, Berlin, Germany.

Guy Emerson and Thierry Declerck. 2014. SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 30–38, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Astrid Erll. 2022. Biographie und Gedächtnis. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 113–122. J.B. Metzler, Stuttgart, Germany.

Katherine Faull. 2021. Visualizing religious networks, movements, and communities: building Moravian Lives. In *Digital Humanities and Christianity*, pages 213–236. De Gruyter, Berlin, Germany.

Katherine Faull and Michael McGuire. 2022. Analyzing Moravian Feelings Using Computational Methods to Ask Questions about Norms and Sentiments in Eighteenth-Century Moravian *Lebensläufe*. *Journal of Moravian History*, 22(2):125–149.

Katherine M. Faull. 1992. The American Lebenslauf: Women's Autobiography in Eighteenth-Century Moravian Bethlehem. *Yearbook of German-American Studies*, 27:23–48.

Katherine M. Faull, editor. 1997. *Moravian women's memoirs: their related lives, 1750 - 1820*, 1 edition. Syracuse University Press, Syracuse, NY, USA.

Jakob Fehle, Thomas Schmidt, and Christian Wolff. 2021. Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques. In *Proceedings of the 17th Conference on Natural Language Processing*, pages 86–103, Düsseldorf, Germany. KONVENS 2021 Organizers.

David Fleischhacker, Roman Kern, and Wolfgang Göderle. 2025. Enhancing OCR in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries*, 26(1):3.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing Biases in Human Perception of User Age and Gender from Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Nophar Geifman, Raphael Cohen, and Eitan Rubin. 2013. Redefining meaningful age groups in the context of disease. *AGE*, 35(6):2357–2366.

Gemma Team. 2025. Gemma 3 technical report.

Janet Z. Giele. 1980. Adulthood as Transcendence of Age and Sex. In Neil J. Smelser, editor, *Themes of work and love in adulthood*, 2 edition, pages 151–172. Harvard University Press, Cambridge, MA, USA.

Janet Z. Giele and Glen H. Jr. Elder. 1998. Life Course Research. Development of a Field. In Janet Z. Giele and Glen H. Jr. Elder, editors, *Methods of Life Course Research. Qualitative and Quantitative Approaches*, pages 5–27. SAGE Publications, Inc., Thousand Oaks, CA, USA.

Evelyn Gius and Michael Vauth. 2022. Towards an Event Based Plot Model. A Computational Narratology Approach. *Journal of Computational Literary Studies*, 1(1).

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 1627–1632, Marseille, France. European Language Resources Association.

Bonnie J. Gunzenhauser. 2001. Autobiography: General Survey. In Margareta Jolly, editor, *Encyclopedia of Life Writing. Autobiographical and Biographical Forms*, volume 1, pages 75–78. Fitzroy Dearborn Publishers, London, UK.

Peter Haehner, Bernd Schaefer, Debora Brickau, Till Kaiser, and Maike Luhmann. 2024. The perception of major life events across the life course. *PLoS ONE*, 19(12):e0314011.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, TX, USA. Association for Computational Linguistics.

Mustapha Hankar, Toufik Mzili, Mohammed Kasri, and Abderrahim Beni-Hssane. 2025. Sentiment analysis survey: datasets, techniques, applications, tools, and challenges. *Knowledge and Information Systems*, 67:8219–8265.

Levke Harders and Hannes Schweiger. 2022. Kollektivbiographische Ansätze. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 285–291. J.B. Metzler, Stuttgart, Germany.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational Linguistics - Volume 1*, pages 299–305, Saarbrücken, Germany. Association for Computational Linguistics.

Robert J. Havighurst. 1976. *Developmental Tasks and Education*, 4 edition. David McKay, New York, NY, USA.

J. Berenike Herrmann and Giulia Grisot. 2022. Lieblingsgegenden, Fenster und Mauern. Zur emotionalen Enkodierung von Raum in Deutschschweizer Prosa zwischen 1850 und 1930. In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, pages 166–170, Potsdam, Germany.

Hans Günter Hockerts. 2008. Vom nationalen Denkmal zum biographischen Portal. Die Geschichte von ADB und NDB 1858–2008. In Lothar Gell, editor, *„... für deutsche Geschichts- und Quellenforschung". 150 Jahre Historische Kommission bei der Bayerischen Akademie der Wissenschaften*, pages 229–269. Oldenbourg, München, Germany.

Michaela Holdenried. 2022. Biographie vs. Autobiographie. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 49–57. J.B. Metzler, Stuttgart, Germany.

Fernando Hsieh, Rafael Dias, and Ivandré Paraboni. 2018. Author Profiling from Facebook Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-Shot Transfer Learning for Event Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Christoph Hube, Maximilian Idahl, and Besnik Fetahu. 2020. Debiasing Word Embeddings from Sentiment Associations in Names. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 259–267, New York, NY, USA. Association for Computing Machinery.

C. J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 216–225.

Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2022. Linked Data – A Paradigm Shift for Publishing and Using Biography Collections on the Semantic Web. In *Proceedings of the Third Conference on Biographical Data in a Digital World*, pages 16–23, Varna, Bulgaria. CEUR Workshop Proceedings.

Peter Hühn. 2014. Event and Eventfulness. In Peter Hühn, Jan Christoph Meister, John Pier, and Wolf Schmid, editors, *Handbook of Narratology*, pages 159–178. De Gruyter, Berlin, Germany.

Marcio Inácio, Gabriela Wick-pedro, and Hugo Goncalo Oliveira. 2023. What do Humor Classifiers Learn? An Attempt to Explain Humor Recognition Models. In *Proceedings of the 7th*

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana, IL, USA.

Matthew L. Jockers. 2015. Revealing Sentiment and Plot Arcs with the Syuzhet Package.

Hyuckchul Jung and Amanda Stent. 2013. ATT1: Temporal Annotation Using Big Windows and Rich Syntactic and Semantic Features. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 20–24, Atlanta, GA, USA. Association for Computational Linguistics.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In *Eurographics Conference on Visualization - State of the Art Reports*, pages 83–103, Cagliari, Italy. Eurographics Association.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Nancy Karweit and David Kertzer. 1998. Data Organization and Conceptualization. In Janet Z. Giele and Glen H. Jr. Elder, editors, *Methods of Life Course Research. Qualitative and Quantitative Approaches*, pages 81–97. SAGE Publications, Inc., Thousand Oaks, CA, USA.

Bettina M. J. Kern, Andreas Baumann, Thomas E. Kolb, Katharina Sekanina, Klaus Hofmann, Tanja Wissik, and Julia Neidhardt. 2021. A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis. In *3rd Conference on Language, Data and Knowledge*, volume 93 of *Open Access Series in Informatics*, pages 37:1–37:17, Zaragoza, Spain. Schloss-Dagstuhl – Leibniz Zentrum für Informatik.

Evgeny Kim and Roman Klinger. 2021. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. Version 2.0. *Zeitschrift für digitale Geisteswissenschaften*.

Jannis Klähn, Janos Borst-Graetz, and Manuel Burghardt. 2025. From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data. *Computational Humanities Research*, 1:e4.

Philipp Koncar, Alexandra Fuchs, Elisabeth Hobisch, Bernhard C. Geiger, Martina Scholger, and Denis Helic. 2020. Text sentiment in the Age of Enlightenment: an analysis of spectator periodicals. *Applied Network Science*, 5(1):1–32.

Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–320, St. Julian's, Malta. Association for Computational Linguistics.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications, Inc., Los Angeles, CA, USA.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association.

Alexander Lasch. 2023. Unterschiede „zwischen uns & den weißen Leuten“. In *Die Herrnhuter Brüdergemeine im 18. und 19. Jahrhundert*, volume 69 of *Arbeiten zur Geschichte des Pietismus*, pages 531–550. Vandenhoeck & Ruprecht.

Alexander Lasch. 2024. *who called them, Sunday *Indians or Shwannaks, that is, white people, the most opprobrious name they could invent. Powerful Constructions in the Service of Verbal Devaluation*, pages 45–70. De Gruyter, Berlin, Germany.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Petri Leskinen, Eero Hyvonen, and Jouni Tuominen. 2018. Analyzing and Visualizing Prosopographical Linked Data Based on Biographies. In *Proceedings of the Second Conference on Biographical Data in a Digital World*, pages 39–44, Linz, Austria. CEUR Workshop Proceedings.

Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2019. *Mining of Massive Datasets*, 3 edition. Cambridge University Press, Cambridge, MA, USA.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bernard Lievegoed. 1981. *Lebenskrisen - Lebenschancen. Die Entwicklung des Menschen zwischen Kindheit und Alter*, 2 edition. Kösel, München, Germany.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18.

Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer US, Boston, MA, USA.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

Christine Lost. 2007. *Das Leben als Lehrtext: Lebensläufe aus der Herrenhuter Brüdergemeine*. Schneider-Verl. Hohengehren, Baltmannsweiler, Germany.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA. Curran Associates Inc.

Lorenzo Lupo, Paul Bose, Mahyar Habibi, Dirk Hovy, and Carlo Schwarz. 2024. DADIT: A Dataset for Demographic Classification of Italian Twitter Users and a Comparison of Prediction

Methods. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4322–4332, Torino, Italia. European Language Resources Association.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Michael McGuire. 2021. *Computational Sentiment Analysis of an 18th Century Corpus of Moravian English Memoirs*. Ph.D. thesis, Indiana University.

Daniel R. Meister. 2018. The biographical turn and the case for historical biography. *History Compass*, 16(1):e12436.

Gisela Mettele. 2007. Constructions of the Religious Self. Moravian Conversion and Transatlantic Communication. *Journal of Moravian History*, 2:7–36.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 321–327, Atlanta, GA, USA. Association for Computational Linguistics.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222.

Franco Moretti. 2005. *Graphs, maps, trees: Abstract Models for a Literary History*. Verso, London, UK.

Marius Mosbach. 2023. Analyzing Pre-trained and Fine-tuned Language Models. In *Proceedings of the Big Picture Workshop*, pages 123–134, Singapore. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Alexander P. D. Mourelatos. 1978. Events, Processes, and States. *Linguistics and Philosophy*, 2(3):415–434.

Alexa Mousley, Richard A. I. Bethlehem, Fang-Cheng Yeh, and Duncan E. Astle. 2025. Topological turning points across the human lifespan. *Nature Communications*, 16(1):10055.

Martin Mueller. 2014. Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment. *Digital Humanities Quarterly*, 8(3).

Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid, and Erik Velldal. 2024. It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments. In *Proceedings of the First Workshop on Patient-Oriented Language Processing*, pages 8–19, Torino, Italia. European Language Resources Association.

Eric T. Nalisnick and Henry S. Baird. 2013. Extracting Sentiment Networks from Shakespeare's Plays. In *2013 12th International Conference on Document Analysis and Recognition*, pages 758–762, Washington, DC, USA. IEEE.

Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. 2011. Author Age Prediction from Text using Linear Regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123, Portland, OR, USA. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

NLP Town. 2023. bert-base-multilingual-uncased-sentiment (revision edd66ab).

Ansgar Nünning. 2022. Fiktionalität, Faktizität, Metafiktion. In Christian Klein, editor, *Handbuch Biographie*, pages 29–35. J.B. Metzler, Stuttgart, Germany.

Juri Opitz. 2024. A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice. *Transactions of the Association for Computational Linguistics*, 12:820–836.

Janis Pagel, Nils Reiter, Ina Rosiger, and Sarah Schulz. 2018. A Unified Text Annotation Workflow for Diverse Goals. In *Proceedings of the Workshop for Annotation in Digital Humanities*, pages 31–36, Sofia, Bulgaria. CEUR Workshop Proceedings.

Alessio Palmero Aprosio and Sara Tonelli. 2015. Recognizing Biographical Sections in Wikipedia. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Lisbon, Portugal. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker and Lori D. Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85:291–301.

Paul Peucker. 2012. Selection and Destruction in Moravian Archives Between 1760 and 1810. *Journal of Moravian History*, 12(2):170–215.

Paul Peucker. 2023. *Herrnhuter Wörterbuch. Kleines Lexikon von brüderischen Begriffen*, 2 edition. Comenius-Buchhandlung, Herrnhut, Germany.

Axel Pichler and Nils Reiter. 2022. From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities. *Journal of Cultural Analytics*, 7(4).

Martin Prell. 2022. Moravians@Sea: A Website for Exploring and Experiencing Moravian Sea Voyages of the Eighteenth Century. *Journal of Moravian History*, 22(2):178–186.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. 2022. Computational Understanding of Narratives: A Survey. *IEEE Access*, 10:101575–101594.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5, 31.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Simone Rebora, Marina Lehmann, Anne Heumann, Wei Ding, and Gerhard Lauer. 2023. Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children's Literature. In *Proceedings of the Computational Humanities Research Conference 2023*, pages 333–343, Paris, France. CEUR Workshop Proceedings.

Simone Rebora, Thomas Messerli, and J. Berenike Herrmann. 2022. Towards a Computational Study of German Book Reviews - A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis. In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, pages 354–356, Potsdam, Germany.

Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. 2017. Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters. In *Proceedings of the Events and Stories in the News Workshop*, pages 42–51, Vancouver, Canada. Association for Computational Linguistics.

Matthias Reinert and Dirk Scholz. 2017. Digital Humanities and the "Deutsche Biographie" as historical biographical information system. In *DH. Opportunities and Risks. Connecting Libraries and Research*, Berlin, Germany.

Matthias Reinert, Maximilian Schrott, and Bernhard Ebneth. 2015. From Biographies to Data Curation – the Making of www.deutsche-biographie.de. In *Proceedings of the First Conference on Biographical Data in a Digital World*, pages 13–19, Amsterdam, the Netherlands. CEUR Workshop Proceedings.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Resources and Evaluation*, pages 1168–1171, Valletta, Malta. European Language Resources Association.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Myriam Isabell Richter and Bernd Hamacher. 2022. Biographische Kleinformen. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 199–205. J.B. Metzler, Stuttgart, Germany.

James C. Riley. 2005. Estimates of Regional and Global Life Expectancy, 1800–2001. *Population and Development Review*, 31(3):537–543.

Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. 2022. Multi-source Multi-domain Sentiment Analysis with BERT-based Models. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 581–589, Marseille, France. European Language Resources Association.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France. PMLR.

Sven Roseen. 1994. *The Dansbury Diaries: Moravian Travel Diaries, 1748-1755, of the Reverend Sven Roseen and Others in the Area of Dansbury, Now Stroudsburg, Pa.* Picton Press.

Kerstin Roth. 2025. *Lebensbeschreibungen im Diskursuniversum Religion: Herrnhutischer Sprachgebrauch im 18. Jahrhundert*. Ph.D. thesis, De Gruyter, Berlin, Germany.

Patrick Sahle. 2016. What is a Scholarly Digital Edition? In Matthew James Driscoll and Elena Pierazzo, editors, *Digital Scholarly Editing. Theories and Practices*, 1 edition, volume 4 of *Digital Humanities*, pages 19–40. Open Book Publishers.

Matthias Schlögl and Katalin Lejtovicz. 2017. A Prosopographical Information System (APIS). In *Proceedings of the Second Conference on Biographical Data in a Digital World*, pages 53–58, Linz, Austria. CEUR Workshop Proceedings.

Pia Schmid. 2004. Frömmigkeitspraxis und Selbstreflexion. Lebensläufe von Frauen der Herrnhuter Brüdergemeinde aus dem 18. Jahrhundert. *Der Bildungsgang des Subjekts. Bildungstheoretische Analysen*, 48:48–57.

Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. 2018. Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. In *Proceedings of the Workshop on Annotation in Digital Humanities co-located with ESSLLI 2018*, pages 47–52, Sofia, Bulgaria. CEUR Workshop Proceedings.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2022. Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800. In *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, pages 107–113, Potsdam, Germany.

Thomas Schmidt, Katrin Dennerlein, Christian Wolff, M. Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, J. Weis, Ulrike Wuttke, Valentin Henning, and Erik Seitz. 2021. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*, Esch-sur-Alzette, Luxembourg. Melusina Press.

Falko Schnicke. 2022a. Begriffsgeschichte: Biographie und verwandte Termini. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 3–9. J.B. Metzler, Stuttgart, Germany.

Falko Schnicke. 2022b. Biographik im 18. Jahrhundert. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 345–354. J.B. Metzler, Stuttgart, Germany.

Astrid Schubring, Jochen Mayer, and Ansgar Thiel. 2019. Drawing Careers: The Value of a Biographical Mapping Method in Qualitative Health Research. *International Journal of Qualitative Methods*, 18:1609406918809303.

Hannes Schweiger. 2022. Biographiewürdigkeit. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 43–48. J.B. Metzler, Stuttgart, Germany.

Mark J. Sciuchetti Jr. 2022. Mapping, Fieldwork, and the Moravian Archives. *Journal of Moravian History*, 22(2):194–205.

Richard A. Settersten and Karl Ulrich Mayer. 1997. The Measurement of Age, Age Structuring, and the Life Course. *Annual Review of Sociology*, 23:233–261.

Prasha Shrestha, Nicolas Rey-Villamizar, Farig Sadeque, Ted Pedersen, Steven Bethard, and Thamar Solorio. 2016. Age and Gender Prediction on Health Forum Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3394–3401, Portorož, Slovenia. European Language Resources Association.

Uladzimir Sidarenka. 2017. PotTS at GermEval-2017 Task B: Document-Level Polarity Detection Using Hand-Crafted SVM and Deep Bidirectional LSTM Network. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 49–54, Berlin, Germany.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Loitongbam Gyanendro Singh and Sanasam Ranbir Singh. 2021. Empirical study of sentiment analysis tools and techniques on societal topics. *Journal of Intelligent Information Systems*, 56(2):379–407.

Noa Visser Solissa, Andreas van Cranenburgh, and Federico Pianzola. 2025. Event Detection between Literary Studies and NLP. A Survey, a Narratological Reflection, and a Case Study. *Journal of Computational Literary Studies*, 4(1).

Rachele Sprugnoli and Sara Tonelli. 2017. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506.

Sophia Stotz, Valentina Stuß, Matthias Reinert, and Maximilian Schrott. 2015. Interpersonal Relations in Biographical Dictionaries. A Case Study. In *Proceedings of the First Conference on Biographical Data in a Digital World*, pages 74–80, Amsterdam, the Netherlands. CEUR Workshop Proceedings.

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.

Marco Antonio Stranisci, Enrico Mensa, Rossana Damiano, Daniele Radicioni, and Ousmane Diakite. 2022. Guidelines and a Corpus for Extracting Biographical Events. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 20–26, Marseille, France. European Language Resources Association.

Debra Taylor. 2001. Obituaries. In Margareta Jolly, editor, *Encyclopedia of Life Writing. Autobiographical and Biographical Forms*, volume 2, pages 666–667. Fitzroy Dearborn Publishers, London, UK.

Karen L. Taylor. 2016. The Discovery of Childhood. In *Childhood through the Looking Glass*, pages 1–13. Brill, Leiden, the Netherlands.

Serge ter Braake and Antske Fokkens. 2015. How to Make it in History. Working Towards a Methodology of Canon Research with Digital Methods. In *Proceedings of the First Conference on Biographical Data in a Digital World*, pages 85–93, Amsterdam, the Netherlands. CEUR Workshop Proceedings.

Hans Thomae. 2002. Psychologische Modelle und Theorien des Lebenslaufs. In Gerd Jüttemann and Hans Thomae, editors, *Persönlichkeit und Entwicklung*, number 113 in Beltz Taschenbuch. Beltz, Weinheim, Germany.

Michael Titzmann. 1996. Zeiterfahrung und Lebenslaufmodelle als theoretischer und historischer Problemkomplex. In *Zeiterfahrung und Lebenslaufmodelle in Literatur und Medien. Akten der Literaturwissenschaftlichen Sektion des 7. Internationalen Kongresses der Deutschen Gesellschaft für Semiotik*, volume 19, 3 of *Kodikas/Code. Ars Semeiotica. Special Issue*, pages 155–164, Tübingen, Germany. Narr.

Karsten Michael Tymann, Matthias Lutz, Patrick Palsbroker, and Carsten Gips. 2019. GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen"*, pages 178–189, Berlin, Germany. CEUR Workshop Proceedings.

Ted Underwood. 2017. A Genealogy of Distant Reading. *Digital Humanities Quarterly*, 11(2).

Shyam Upadhyay, Christos Christodoulopoulos, and Dan Roth. 2016. "Making the News": Identifying Noteworthy Events in News Articles. In *Proceedings of the Fourth Workshop on Events*, pages 1–7, San Diego, CA, USA. Association for Computational Linguistics.

Wouter Van Atteveldt, Mariken A. C. G. Van Der Velden, and Mark Boukes. 2021. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2):121–140.

Arnold Van Gennep. 1960. *The rites of passage*. University of Chicago Press, Chicago, IL, USA.

Jacqueline Van Gent. 2017. Moravian Memoirs and the Emotional Salience of Conversion Rituals. In Merridee L. Bailey and Katie Barclay, editors, *Emotion, Ritual and Power in Europe, 1200–1920*, Palgrave Studies in the History of Emotion, pages 241–260. Palgrave Macmillan.

David L. Vander Meulen and G. Thomas Tanselle. 1999. A System of Manuscript Transcription. *Studies in Bibliography*, 52:201–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated Event Annotation in Literary Texts. In *Proceedings of the Conference on Computational Humanities Research*, pages 333–345, Amsterdam, the Netherlands. CEUR Workshop Proceedings.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore. Association for Computational Linguistics.

Kurt Vonnegut. 1981. *Palm Sunday: an autobiographical collage*. Delacorte Press, New York, NY, USA.

Melissa L. H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, 41(2):534–538.

Uli Waltinger. 2010a. Sentiment Analysis Reloaded. A Comparative Study on Sentiment Polarity Identification Combining Machine Learning and Subjectivity Features. In *Proceedings of the 6th International Conference on Web Information Systems and Technology*, pages 203–210, Valencia, Spain. SciTePress - Science and and Technology Publications.

Ulli Waltinger. 2010b. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.

Zijian Wang, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067, New York, NY, USA. Association for Computing Machinery.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Lukas Werner. 2022. Deutschsprachige Biographik. In Christian Klein, editor, *Handbuch Biographie: Methoden, Traditionen, Theorien*, pages 395–409. J.B. Metzler, Stuttgart, Germany.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2018. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany. DuEPublico.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A Survey of Event Extraction From Text. *IEEE Access*, 7:173111–173137.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, 50(2):25:1–25:33.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3912–3921, Hong Kong, China. Association for Computational Linguistics.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of Happy Endings in German Novels based on Sentiment Information. In *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing*, pages 9–16, Riva del Garda, Italy. CEUR Workshop Proceedings.

Albin Zehe, Martin Becker, Fotis Jannidis, and Andreas Hotho. 2017. Towards Sentiment Analysis on German Literature. In Gabriele Kern-Isberner, Johannes Fürnkranz, and Matthias Thimm, editors, *KI 2017: Advances in Artificial Intelligence*, volume 10505 of *Lecture Notes in Computer Science*, pages 387–394. Springer, Cham, Switzerland.

Kai Zhang, Qiuxia Zhang, Chung-Che Wang, and Jyh-Shing Roger Jang. 2025. Comprehensive Study on Zero-Shot Text Classification Using Category Mapping. *IEEE Access*, 13:23526–23546.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Andrea Zielinski, Calvin Spolwind, Henning Kroll, and Anna Grimm. 2023. A Dataset for Explainable Sentiment Analysis in the German Automotive Industry. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 138–148, Toronto, Canada. Association for Computational Linguistics.

Emrullah Şahin, Naciye Nur Arslan, and Durmuş Özdemir. 2025. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 37(2):859–965.

Appendix A

Class definitions

A.1 Sentiment classes

Label	Definition
`negative`	Use this label whenever the final state or result of the action or event described in a given sentence is depicted in a (rather) negative way.
`neutral`	Use this label whenever all actions or events described in the given sentence are depicted without a (clear) polarity or whenever the given sentence lacks any sentiment-related information.
`positive`	Use this label whenever the final state or result of the action or event described in a given sentence is depicted in a (rather) positive way.

Table A.1: Definitions of sentiment labels

A.2 Life stage classes

Label	Definition
`early childhood`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 0–5 years old.
`childhood`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 6–12 years old.
`adolescence`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 13–17 years old.
`early childhood`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 18–29 years old.
`adulthood`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 30–59 years old.
`old-age`	Use this label whenever the main event described in the given sentence occurred when the biographical subject was 60 years old or older.
`NONE`	Use this label whenever the given sentence is a non-biographical remark of the author.

Table A.2: Definitions of life stage labels

A.3 Life event domain classes

Label	Definition
birth	Use this label whenever the main event described in the given sentence is the birth of the biographical subject.
collective	Use this label whenever the main event described in the given sentence affects a whole population. This includes e.g. war, pandemic, climate crisis.
death	Use this label whenever the main event described in the given sentence is the death of the biographical subject.
education	Use this label whenever the main event described in the given sentence prepares the career of the biographical subject. This includes e.g. training, further education, studies, degrees.
family	Use this label whenever the main event described in the given sentence is related to relatives of the biographical subject. This includes e.g. their birth, death, marriage, separation, or conflicts or reconciliations within the family.
health	Use this label whenever the main event described in the given sentence is related to the health of the biographical subject. This includes e.g. injuries, illnesses, recovery, hospitalization, diagnoses.
housing	Use this label whenever the main event described in the given sentence is related to the personal housing situation of the biographical subject. This includes e.g. moving, loss of a flat or house, alterations, renovations.
legal	Use this label whenever the main event described in the given sentence is related to legal issues of the biographical subject. This includes e.g. involvement in legal disputes, facing legal consequences, engaging with legal processes.
leisure	Use this label whenever the main event described in the given sentence is related to personal leisure activities of the biographical subject. This includes e.g. holidays, leisure activities like swimming or hiking, purchases for leisure activities, relaxation and recreation, hobbies.
money	Use this label whenever the main event described in the given sentence is related to financial matters of the biographical subject. This includes e.g. personal insolvency, inheritance, debts, borrowing, taxation.
other_social	Use this label whenever the main event described in the given sentence is related to friends, neighbors, or acquaintances of the biographical subject.
religious	Use this label whenever the main event described in the given sentence is either a significant religious or spiritual event or ceremony the biographical subject participated in, or a milestone in her or his faith or belief system.
romantic	Use this label whenever the main event described in the given sentence is related to partnership developments of the biographical subject. This includes e.g. wedding, engagement or marriage proposal, separation, a new partner, other relationship events.
work	Use this label whenever the main event described in the given sentence is related to the occupation of the biographical subject. This includes e.g. new professional activities, job loss, changes in the context of the own career.
no_or_other	Use this label for all cases not covered with one of the other labels.

Table A.3: Definitions of all 15 life event domain labels used for the manual annotations

Label	Definition
birth	Use this label whenever the main event described in the given sentence is the birth of the biographical subject.
death	Use this label whenever the main event described in the given sentence is the death of the biographical subject.
education_or_work	Use this label whenever the main event described in the given sentence either prepares the career of the biographical subject or is related to her or his occupation. This includes e.g. training, further education, studies, degrees, new professional activities, job loss, changes in the context of the own career.
family	Use this label whenever the main event described in the given sentence is related to relatives of the biographical subject. This includes e.g. their birth, death, marriage, separation, or conflicts or reconciliations within the family.
health	Use this label whenever the main event described in the given sentence is related to the health of the biographical subject. This includes e.g. injuries, illnesses, recovery, hospitalization, diagnoses.
housing	Use this label whenever the main event described in the given sentence is related to the personal housing situation of the biographical subject. This includes e.g. moving, loss of a flat or house, alterations, renovations.
leisure	Use this label whenever the main event described in the given sentence is related to personal leisure activities of the biographical subject. This includes e.g. holidays, leisure activities like swimming or hiking, purchases for leisure activities, relaxation and recreation, hobbies.
religious	Use this label whenever the main event described in the given sentence is either a significant religious or spiritual event or ceremony the biographical subject participated in, or a milestone in her or his faith or belief system.
romantic	Use this label whenever the main event described in the given sentence is related to partnership developments of the biographical subject. This includes e.g. wedding, engagement or marriage proposal, separation, a new partner, other relationship events.
no_or_other	Use this label for all cases not covered with one of the other labels.

Table A.4: Definitions of the reduced set of ten life event domain labels

Appendix B

Sentiment polarity of life event domains

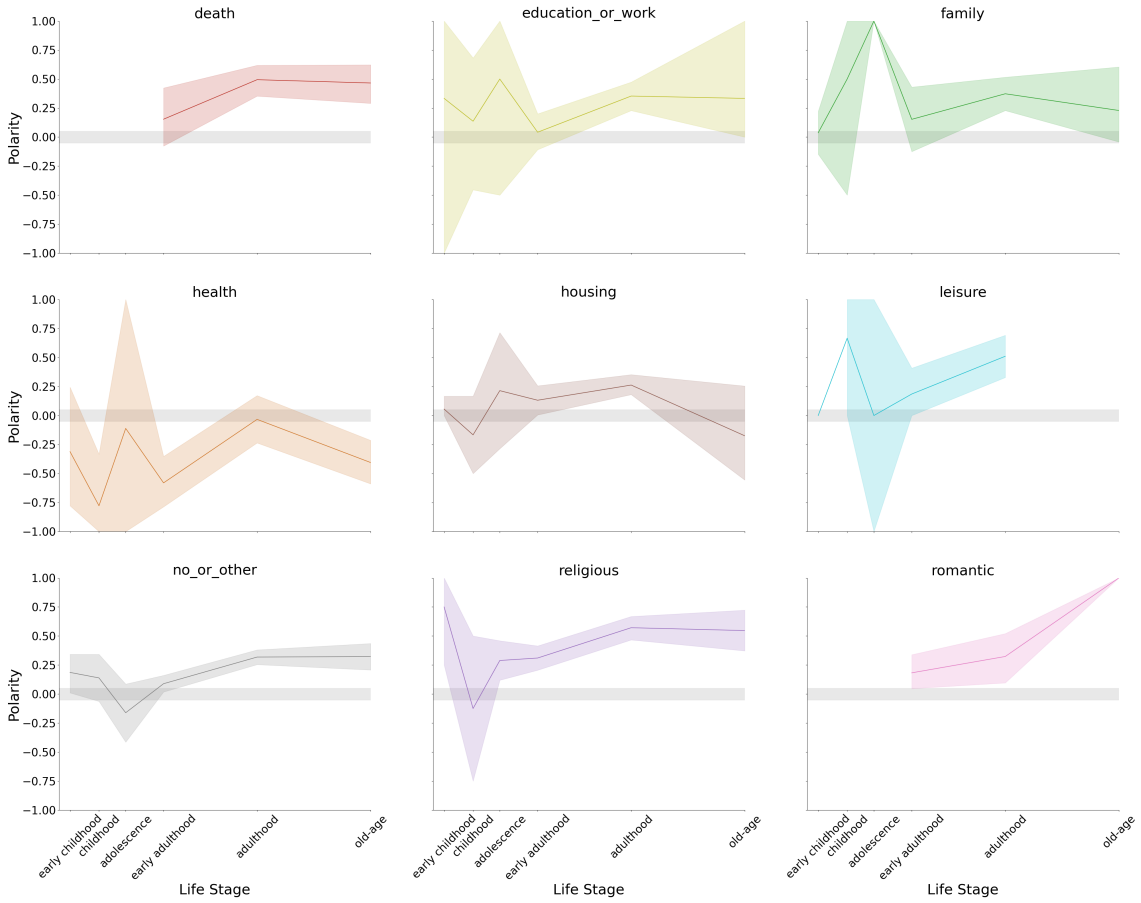

Figure B.1: Sentiment polarity of life event domains in 89 Moravian memoirs over narrated time as predicted by gbert*

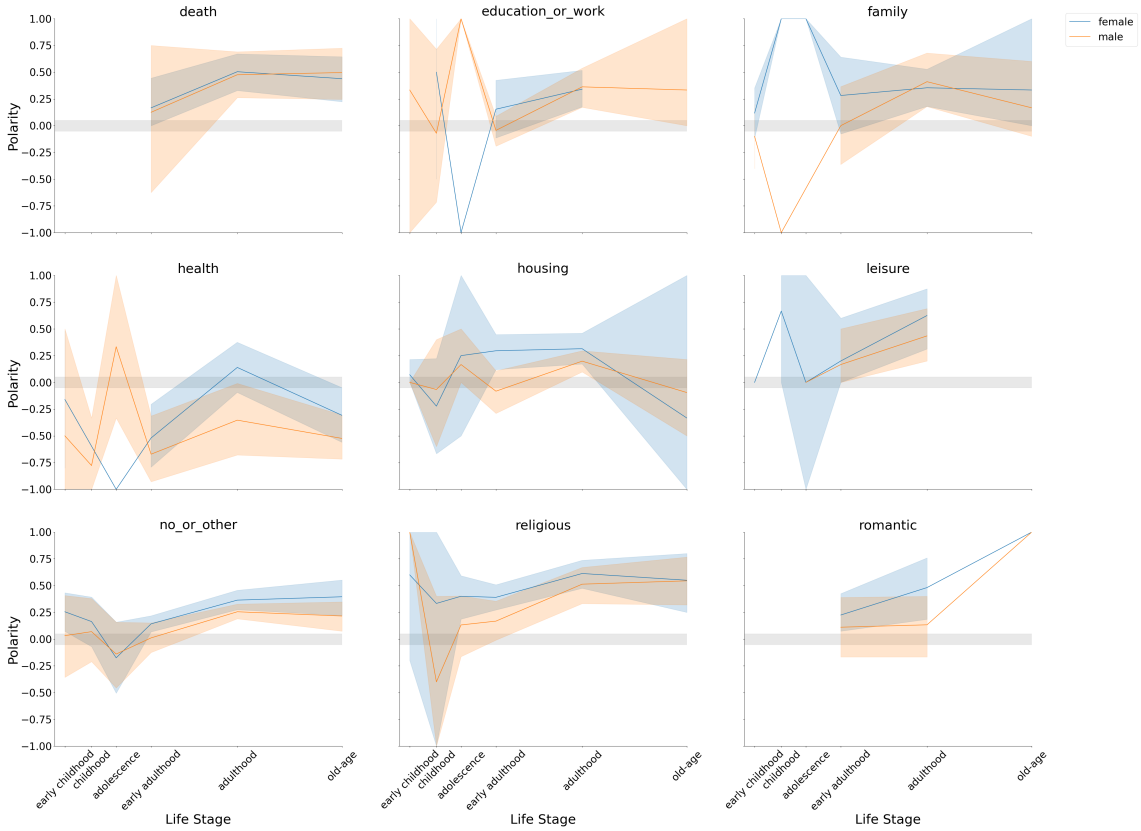

Figure B.2: Sentiment polarity of life event domains in 89 Moravian memoirs over narrated time grouped by subject gender as predicted by gbert*

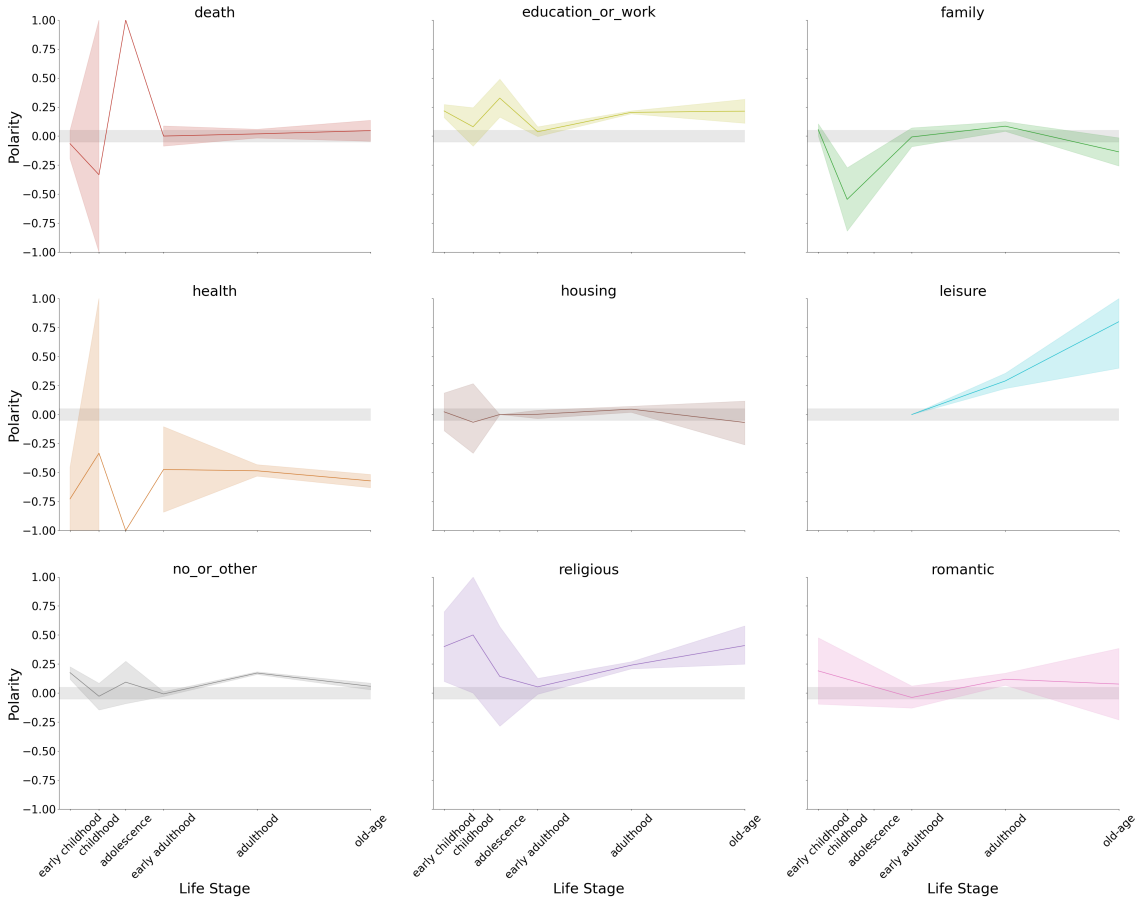

Figure B.3: Sentiment polarity of life event domains in 2630 ADB entries over narrated time grouped by subject gender as predicted by gbert*

Appendix C

Runtimes per System

System	System Type	Training	Inference		
			Test	MM	ADB
AffectiveNorms	dictionary	–	00:01	–	–
BAWL-R	dictionary	–	00:00	–	–
Germanlex	dictionary	–	00:00	–	–
GermanPolarityClues	dictionary	–	00:00	00:00	–
MultilingualSentiment_de	dictionary	–	00:00	–	–
SentiMerge	dictionary	–	00:00	–	–
SentiWS	dictionary	–	00:00	–	–
GerVADER	dictionary + rules	–	00:00	00:00	–
SVM*	SVM*	00:04	00:01	00:03	02:40
bert*	BERT*	*08:45*	01:10	–	–
bert_multilingual*	BERT*	*09:47*	00:58	–	–
gbert*	BERT*	*08:55*	00:45	03:50	02:55:27
german_sentiment_bert	BERT	–	00:50	–	–
german_sentiment_bert*	BERT*	*08:25*	00:45	–	–
multilingual_sentiment_bert	BERT	–	00:30	–	–
multilingual_sentiment_bert*	BERT*	*09:34*	00:50	–	–
bert_110M	zero-shot (BERT)	–	03:53	–	–
bert_multilingual_110M	zero-shot (BERT)	–	06:20	–	–
gbert_110M	zero-shot (BERT)	–	06:47	–	–
bart-nli_406M	zero-shot (NLI)	–	34:10	–	–
mDeBERTa-nli_279M	zero-shot (NLI)	–	14:15	–	–
Gemma3_27B	zero-shot (LLM)	–	*02:20*	–	–
Nemotron_Ultra_253B	zero-shot (LLM)	–	*02:51*	*10:19*	–

Table C.1: Mean runtimes per system for training with 1768, testing with 442, and inference on 2210 Moravian and 115,486 ADB sentences (in hours, minutes and seconds in the form (hh:)mm:ss; GPU usage in italics)