Predictive Analytics on Emotional Data Mined from Digital Social Networks with a Focus on Financial Markets

Inaugural Dissertation

for the

Attainment of the Doctorate

from the

Faculty of Management, Economics and Social Sciences

at the

University of Cologne

2019

presented

by

Dipl.-Wirt.-Inf. Stefan Nann

from

Bonn

Lecturer: Prof. Dr. Detlef Schoder

Co-Lecturer: Prof. Dr. Jörn Grahl

Day of Dissertation: April 22nd, 2020

Table of Contents

Figure Index

Table Index

Abbreviation Index

UGC:	User-Generated Content
SNA:	Social Network Analysis
API:	Application Programming Interface
DJIA:	Dow Jones Industrial Average
S&P 500:	Standard & Poor's 500 index
LDA:	Latent Dirichlet Allocation
IS:	Information Systems
SMA:	Simple Moving Average
ROI:	Return on Investment

1. Introduction

André Kostolany, a successful stock market investor, made the well-known remark that facts account for only 10 percent of the reactions on the stock market; everything else is seemingly driven by psychology. He believed strongly in this investment philosophy and became one of the most renowned and successful investors of the 20th century. He was an early anticipator of the great Stock Market Crash of 1929 and maintained his successful investment career for nearly 70 years. Despite that the Internet already existed when Kostolany died in 1999, objective data about his investments were not the key to his success. His decision-making process most likely depended, to a large part, on intuition and hard-earned experience.

André Kostolany was not the only person interested in the psychology of markets in earlier days. The possible influence of mood and sentiment on financial markets also caught the attention of John Maynard Keynes (1936). The intuition of both men became measurable with the rise of the Internet and new technologies that evolved with it. One of these technologies is "social media," generally comprised of online spaces in which people exchange information and data digitally. In the sections that follow, the material of that exchange is referred to as User-Generated Content (UGC)[1] as a more general representation for social media content.

Social media and UGC are examples of data, the 21st century's most important resource. As British mathematician Clive Humby put it in 2006, coining a new phrase, "Data is the new oil." We live today in a data-driven economy. Data are being created all the time without us even noticing it. Much of what we do each day now takes place in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. According to the Global Web Index, individual Internet users have 7.6 social media accounts, on average.[2] We Are Social Ltd and Hootsuite Inc. state that social media users grew by 320 million between September 2017 and October 2018. In a 2011 study, AOL/Nielsen showed that 27 million pieces of content were shared every day in social media.[3]

Using mobile devices, shopping online, reading online news and blogs, employing GPS services, using voice control systems, and optimizing every part of our lives with apps is now typical for almost all of us. A larger consumption of data also means a larger provision of our own data. People like to share data in social media, whether they are consumed from external sources or, in many cases, personal data. Often, people are affected by emotions based on the data they consume (e.g., Guillory et al. 2011, Hancock et al. 2008, Kramer et al. 2014). Sharing these particular data creates new emotions for other

[1] The working definition of UGC for this thesis is the definition given in the OECD report by Vickery and Wunsch-Vincent (2007, p. 18).
[2] https://blog.globalwebindex.com/chart-of-the-day/internet-users-have-average-of-7-social-accounts/
[3] https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/

people (e.g., Kramer 2012), creating a cycle in which data create emotions and new data are generated by sharing these emotions – for example, through social media or UGC.

Financial markets are especially dependent on data. Having information in advance can increase the return on investment that is realized. For example, a stock's price is often viewed as the result of all information available at a certain point in time, with new information being priced in rapidly as it becomes available. Communication behavior in the financial industry, as in many other industries, has changed radically in recent years, driven by technological developments. Related information and data are, to a large extent, shared via the Internet and, eventually, through social media. Since financial markets are characterized by fast-flowing information, "emotional data" extracted from these types of sources seem to have a high potential predictive value.

There are many examples of Twitter[4] being used to broadcast information that triggers widespread emotions almost instantly. A recent and quite prominent example comes from Elon Musk, founder and CEO of Tesla Inc. On August 7, 2018, Musk tweeted that he was "considering taking Tesla private at $420. Funding secured." At the time, Tesla's stock was trading at around US$370. Musk's tweet received a lot of immediate attention. People who follow him on Twitter re-tweeted his original tweet and numerous news outlets published the message on their websites. Figure 1 provides a detailed description of the example.

In the Musk example, a simple tweet created a high degree of attention in a very short period. After quick initial reactions, many people – stock market participants in particular – were discussing the potential impact of Musk's tweet on Tesla's stock price. Most of the discussion was speculative given the lack of objective information. Hence, the content of the discussion was mostly subjective and emotional opinion, and the reaction of Tesla's stock price was strong and immediate, with increases in both volatility and trading volume. The NASDAQ market ended up halting trading of Tesla stock for more than an hour – a very rare event.

After a short and sharp increase of the stock price very soon after the tweet, the price dropped in the following days, as most people anticipated that Elon Musk could not simply take his company off the stock market. Later, the *Financial Times* reported the real reason behind Musk's tweet.[5] Regardless of that reason, though, the example illustrates the tremendous impact a single bit of information in social media can have on the stock market.

[4] Twitter (https://www.twitter.com) was founded in 2006 and has experienced exponential growth over the past few years with respect to the number of users and published messages. The latter, called "tweets," are short text messages (maximum 140 characters in length at first; now 280 characters). In 2018, about 500 million tweets were sent per day. Beginning around 2008, Twitter also became relevant to the financial market as users tweeted opinions about stock prices and the market situation.

[5] Financial Times: "Saudi Arabia's sovereign fund builds $2bn Tesla stake", https://www.ft.com/content/42ca6c42-a79e-11e8-926a-7342fe5e173f, downloaded on March 11th, 2019

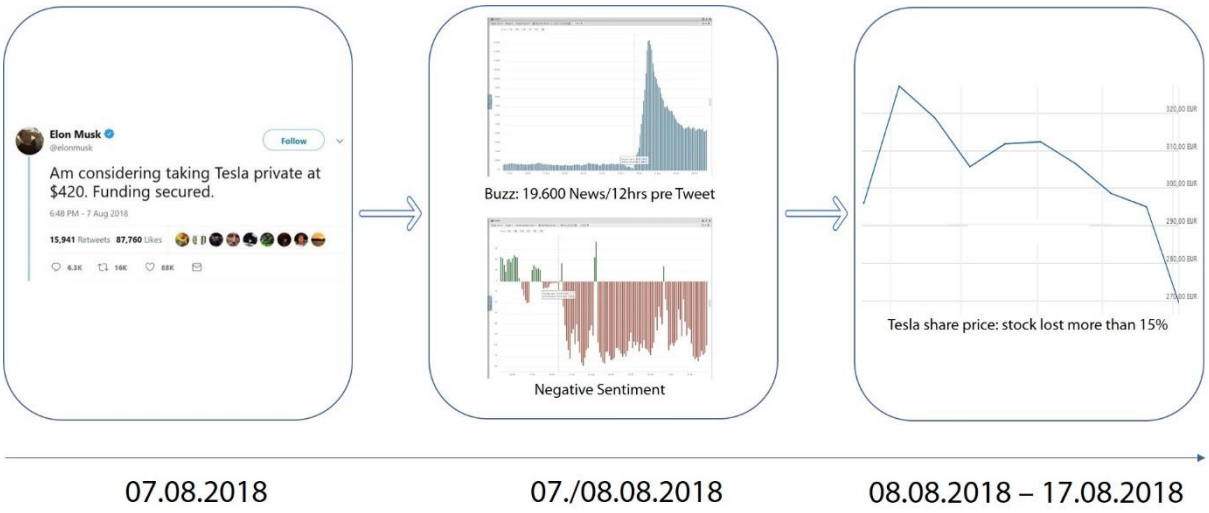

Figure 1: Tweet by Elon Musk: development of message volume and stock price

As the Tesla example shows, people's emotions can influence their decisions. Over the last 30 years, behavioral finance researchers have shown that financial markets are influenced by investor emotions and psychology. Nofer and Hinz (2015) showed that the mood levels of social media participants can predict share returns. Well more than a decade earlier, Daniel et al. (2002) provided an extensive overview of research about the impact of psychological effects on the stock market. The movement of a stock price is an example of those effects, showing that emotions shared via social media can have real-world effects (e.g., Nann et al. 2013).

This fact was a major motivation for the studies presented here and was a primary influence on the research conducted in some of the individual contributions of this dissertation.

There are many more examples that illustrate how significant new digital forms of communication and connection via social media can be for real-world events. Sources are not limited to Twitter; other social media or social network platforms may have similar effects. It is possible to draw conclusions based on the network structures of the users of social media platforms, as in the case of the German digital community Xing, which connects peers and companies and provides a large digital social network (Nann et al. 2009, Contribution III). Ubiquitous global digital networks and a continuous exchange of data make the spread of news and tweets possible at a speed unimaginable ten years ago. As a result, huge amounts of publicly available data offer new possibilities for evaluation and interpretation that go far beyond the possibilities that existed during the 20th century and during the earliest part of this century.

The term "Data Intelligence" has become increasing prominent in this context. Data mining technologies and the increasing power of today's computers support analysis of at least portions of the growing flood of data in real time. Academics and practitioners want to know how much meaningful information can be derived through algorithmic analyses and how the predictive value of this information can be determined automatically to promote better decision making. The next section discusses "Data Intelligence" and a framework for "Emotional Data Intelligence" in more detail.

We will never know, but André Kostolany and John Maynard Keynes may well have dreamt of such methods to measure the psychology of the financial markets. It was their conviction that stock markets depend strongly on psychological effects, mood states, and people's emotions. Perhaps they asked themselves how it might be possible to capture and analyze those emotions and psychology. We do know, however, that social media, UGC, and huge amounts of (digital) data have combined to make it actually possible.

Both academics and practitioners have paid considerable attention to data accumulated on social media platforms in the last ten to fifteen years. Empirical tests of André Kostolany's intuition became possible at the beginning of the current century, when social media facilitated the large-scale collection of emotion-based and sentiment-related data suitable for financial modeling. At the same time, it became possible to quantify and test phenomena from other business fields with the newly available data in social media (e.g., Egger and Lang 2013).

2. A Framework for Emotional Data Intelligence

A comprehensive search yields no relevant academic studies that discuss “Data Intelligence” or “Emotional Data Intelligence” with respect to information system (IS) topics.

Some online sources do, however, provide definitions of “Data Intelligence.” Techopedia, for example states: “Data intelligence is the analysis of various forms of data in such a way that it can be used by companies to expand their services or investments. Data intelligence can also refer to companies' use of internal data to analyze their own operations or workforce to make better decisions in the future. Business performance, data mining, online analytics, and event processing are all types of data that companies gather and use for data intelligence purposes” (Techopedia).

The following paragraphs offer working definitions for this dissertation of both “Data Intelligence” and “Emotional Data Intelligence.”

“Data Intelligence” refers to the collection and analysis of large amounts of data to uncover relationship between and among different data points in a meaningful way for the purpose of promoting better decision making. The methods and tools of “Data Intelligence” focus on understanding data, uncovering alternative explanations, resolving issues, and identifying future trends to improve decisions. User-Generated Content (UGC) comprises a substantial part of communication via social media. In this dissertation, UGC that carries and facilitates the exchange of emotions is referred to as ”emotional data.” When the methods and tools of “Data Intelligence” are applied to data that transport emotions, it can be referred to as “Emotional Data Intelligence.”

People ”produce” emotional data, that is, they express their emotions via tweets, forum posts, blogs, and so on, or they ”consume” it by being influenced by sentiment, feelings, opinions, and the like expressed by others. People are often affected by emotions based on the data they consume (e.g., Guillory et al. 2011, Hancock et al. 2008, Kramer et al. 2014). As stated earlier, sharing these data creates new emotions for different people (e.g., Kramer 2012), creating a cycle in which data create emotions and new data are generated by sharing these emotions. Decisions often depend on these shared emotions and data – which again lead to new data because decisions may change behaviors or results. “Emotional Data Intelligence” ultimately seeks an answer to the question of how all the different emotions expressed in public online sources influence decision-making processes.

Figure 2 introduces a theoretical framework for the interplay of data, society, financial markets, and the interaction between emotions and decision-making processes in the financial markets domain. Of course, the production and consumption of UGC is not limited to financial markets, but is typical in many domains. Thus, the “Financial Markets” row in Figure 2 could easily be substituted with rows devoted to domains such as e-commerce, social networking sites, and so on.

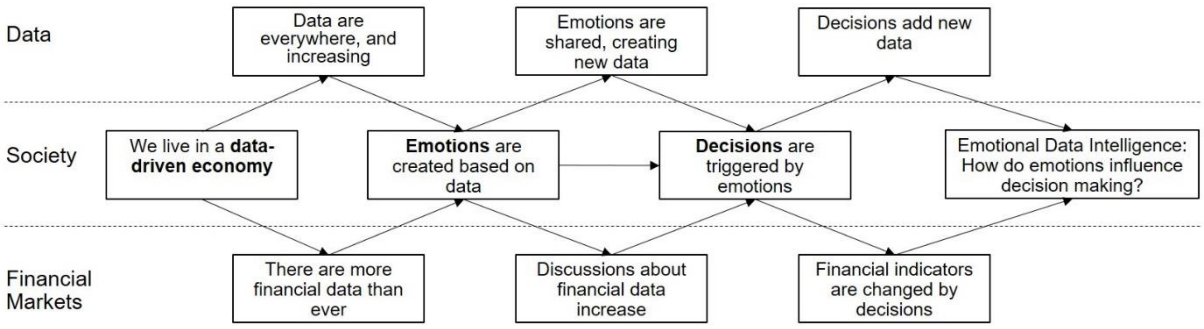

Figure 2: Framework for Emotional Data Intelligence

3. Definition of the Research Scope

Signal or noise: Do network structures and emotional sentiment data extracted from digital social networks contain predictive information, or are they just noise?

This question outlines the main research topic of this dissertation and serves as the primary question for all contributions of this dissertation. It is also the central question that guides a considerable amount of work in recent computational analysis and data mining. There is a unifying theme of studies that examine whether sentiment indicators help in forecasting returns and volatility in global financial markets. Sentiment indicators are often appealing to investors since financial markets are known to be influenced not only by economic fundamentals but also by emotions. The concept "behavioral finance" is often used to explain the mood effect on the stock market (e.g., Daniel et al. 2002, Nofer and Hinz 2015). Behavioral finance and investor sentiment theory have established that the behavior of financial market participants can be shaped by whether they feel optimistic (bullish) or pessimistic (bearish) about future market values (Bollen and Mao 2011). The work of Antweiler and Frank (2004) is a key text about investor sentiment and a combination of sentiment analysis and predictive analytics theory. It has also been the basis for the stock market-related articles of this dissertation. Three contributions – Gloor et al. (2009), Nann et al. (2013), and Janetzko et al. (2017) – use different combinations of network metrics, sentiment analysis methodologies, and predictive analytics approaches to study whether social media and sentiment indicators help explain future developments of prices on the stock market.

Social Network Analysis (SNA) is another important research theory of this dissertation. SNA is also the oldest research field compared to the other research streams of this dissertation (see also section 7.3). Bavelas (1950), Bavelas and Barrett (1951), and Leavitt (1951) introduced important elements of centrality in network analysis. Freeman (1979) and Wasserman and Faust (1994) also provided detailed discussions of the concept of betweenness centrality. Centrality measures and related questions of network analysis are central components of some research contributions of this dissertation. Krauss et al. (2008) studied the influence of social network metrics and sentiment analysis on movie success. Gloor et al. (2009) provided a general model based on SNA to predict trends in different domains based on sentiment analysis and the computation of network centrality measures. Using methods from SNA, Nann et al. (2009) investigated whether online social networking structures can be used to predict entrepreneurial success.

4. Literature Overview

This literature review is organized into two parts. The first part categorizes research related to sentiment analysis, predictive power/analytics of UGC, and SNA. The focus is on literature that forms the basis of the individual contributions of this dissertation. The second part aims to draw as complete a picture as possible of the literature in the field during the last 15 years. While not every paper is discussed in detail, the presentation of the papers in chronological order (clustered by years) and showing how the content of studies was adapted over time aims to demonstrate the steady and rapid development of the field. (Studies published after the papers that comprise this dissertation are summarized in the second part.)

4.1. Main Literature of Individual Contributions of this Dissertation

Sentiment analysis is a broad research field applied in many different domains (Berger et al. 1996, Pang et al. 2002, Whitelaw et al. 2005, Abbasi et al. 2008, Boiy and Moens 2009, Choi et al. 2009, Lin and He 2009, Narayanan et al. 2009, Mizumoto et al. 2012, Fang et al. 2012, Serrano-Guerrero et al. 2015). The predictive power/analytics of UGC research stream is much broader and also applies to many different domains. Chintagunta et al. (2010) and Dellarocas et al. (2007), for instance, found a relationship between online consumer reviews and movie success. Zhu and Zhang (2010) examined the prediction of video game sales. Heimbach and Hinz (2012) looked at music sales. Chevalier and Mayzlin (2006) explored book sales.

Both research streams are closely related: large-scale, automated analysis of UGC is possible only with sophisticated sentiment analysis approaches. Therefore, many studies consider both research streams. Antweiler and Frank (2004), Das and Chen (2007), Bollen et al. (2011), Bollen and Mao (2011), Mao et al. (2011), Zhang and Swanson (2010), and Sprenger and Welpe (2010) are but a few examples of early works that examine the relationship of sentiment, UGC, and stock returns. Other studies related to these research streams evaluate the potential of traditional and editorial news to help predict future stock returns. Cohen and Frazzini (2008), Schumaker and Chen (2009), and Dion et al. (2011) are a few examples of papers in this particular area.

An extensive literature research reveals that the earliest publication about using UGC for financial forecasting is a 2001 study by Tumarkin and Whitelaw. The authors examined the possible link between sentiment and posting volume recorded on a major financial message board. They were also searching for correlations with stock returns and trading volume. They found that message board activity and stock returns were related on days of abnormal message board activity, a key insight. However, the data did not indicate significant predictive value for stock returns or trading volume based on message board activity or sentiment; rather, the findings suggested that developments on the stock market prompted discussions on message boards.

The work of Antweiler and Frank in 2004 is a major milestone in this research field. The study, published in the *Journal of Finance*, showed the relevance of analyzing UGC for the financial market. Similar to

Tumarkin and Whitelaw (2001), Antweiler and Frank (2004) investigated whether Internet message board activity or sentiment is predictive for stock returns or market volatility. They analyzed 1.5 million messages from Yahoo! Finance and Raging Bull, at the time the largest dataset for such analysis, and found that stock messages help predict volatility on stock markets. The effect of stock messages on stock returns was statistically significant but economically small.

A number of other studies have also found predictive evidence of UGC on stock returns (Bagnoli et al. 1999, Tumarkin and Whitelaw 2001, Jones 2006, Gu et al. 2006, Das and Chen 2007, Sabherwal et al. 2008).

Since its establishment in 2006, Twitter has captured a high level of interest in the academic world. It has turned out to be a very valuable source for large-scale, data-driven empirical studies that were never before possible. Twitter also provides fairly simple access to its data via a public application programming interface (API). Straightforward access and public data are surely major reasons that so many academic studies in the years since Twitter's founding have focused on analyzing the real-world effects of tweets. Asur and Huberman (2010), for instance, investigated the predictive power of tweets for box office returns, analyzing 2.89 million tweets related to the topic. Twitter has since been widely used to measure the effect of emotion and mood on the development of stock prices.

Bollen et al. (2011) can be seen as a "first-mover" in analyzing a large-scale dataset of tweets with a focus on the stock market. Their work suggests that some public mood states expressed on Twitter are predictive of changes in Dow Jones Industrial Average (DJIA) closing values. In another study, Bollen and Mao (2011) used Twitter for market prediction and public mood analysis. Mao et al. (2011) compared the value for financial prediction of a range of online datasets (Twitter feeds, news headlines, and volumes of Google search queries). Like earlier researchers, they focused on indices and volatility measures such as the VIX (U.S. market volatility index). While they found that Google search queries and Twitter investor sentiment have predictive value, they identified only lagging indicators of the financial markets for traditional surveys.

Oh and Sheng (2011) examined approximately 72,000 microblog postings from Stocktwits (https://stocktwits.com: a Twitter-like social media platform where users post about financial markets in particular), extending over a three-month period, to predict stock price movements. Applying sentiment analysis, they found microblog messages predicted future stock price movement. They also briefly evaluated potential return on investments, finding that simple (not adjusted returns) delivered better results than market-adjusted returns. Sprenger and Welpe (2010) and later Sprenger et al. (2014) found that stock microblogs may contain alpha information, that is, information that is not yet reflected in price levels on asset markets, thus leading to future movements in line with the efficient market hypothesis (Malkiel and Fama 1970). Rao and Srivastava (2012) used financially relevant tweets and Google search volume to predict stock returns, trading volume, and volatility of different asset classes.

While research in the pre-Twitter era found that discussion mostly follows market movements (e.g., Bagnoli et al. 1999, Tumarkin and Whitelaw 2001, Das and Chen 2007), many studies based on large-scale datasets comprised of tweets have clearly detected predictive value in online data for stock price changes, for instance Bollen et al. (2011), Mao et al. (2011), Sprenger et al. (2014), Zhang and Swanson (2010), Oh and Sheng (2011), or Xu et al. (2012), Nofer and Hinz (2015).

Until 2013, most studies focused on a single source and short time series to build prediction models (e.g., message boards, blogs, editorial news, or Twitter). Nann et al. (2013), which is also a contribution of this dissertation, extends this earlier research by aggregating data from multiple sources (Twitter, several online message boards, and traditional news). Furthermore, historical data over a six-month period were considered, one of the longest time periods and granularity analyzed compared to research that preceded this work. The subjects of prediction were daily stock price movements from the Standard & Poor's 500 index (S&P 500). Most research concerned with stock market predictions based on UGC to that point had been mainly theoretical in nature and did not take into account real-world limitations such as broker fees, bid/ask differences, and liquidity. To demonstrate the potential practical application of the findings of the contribution, a description of a simple trading model based on the predictor, considering commission fees, transaction costs, and stock liquidity was given.

Compared to stock market predictions, only a relatively small number of studies make use of sentiment-based indicators to predict exchange rate movements. Work by Papaioannou et al. (2013) and Janetzko (2014) on predictive modeling of the EUR/USD exchange rate and a study by Crone and Koeppel (2014) that harnessed sentiment indicators as explanatory variables to predict the AUD/USD currency pair provided evidence of sentiment indicators' potential for correct predictions in more than 50 percent of out-of-sample tests. Twitter has also been used to extract sentiment and building prediction models for commodity markets and currency rates (Rao and Srivastava 2012).

Janetzko et al. (2017), the most recent contribution of this dissertation, was motivated by the lack of research into the relationship between UGC and the exchange rate market, the need for comparability between sentiment sources, and by questions of robustness and varying predictive value across different sources. Whereas the majority of previous research was concerned with stocks and indices, this paper contributed by studying sentiment-based predictions for currencies.

SNA and applications of network centrality measures have been used to identify trends before they become recognized by the rest of the world (Kleinberg 2008). Dodge and Kitchin (2000) and Dodge and Kitchin (2002) introduced systems for the static visualization and analysis of the link structure of the Web. In a related stream of work, researchers sought to predict the hidden linking structure based on known links (e.g., Adar et al. 2004, Al Hasan et al. 2006). In Gloor et al. (2009), a contribution of this dissertation, research focused on a similar application – tracking the strengths of concepts over time. The authors used a tool called Condor to conduct SNA-related analysis (Gloor and Zhao 2004).

Extensive network analysis research has also investigated the effect of network structures on the performance of individuals (e.g., Ahuja et al. 2003, Bulkley and Van Alstyne 2006, Cross and Cummings 2003, Gloor et al. 2010, Moran 2005, Sparrowe et al. 2001), groups (Balkundi and Harrison 2006, Brass 1985, Mayo and Pastor 2005, Reagans and Zuckerman 2001, Sparrowe et al. 2001) and organizations (Ahuja 2000, Podolny and Baron 1997, Powell et al. 1996, Raz and Gloor 2007, Uzzi 1996). Mayer and Puller (2008) analyzed friendship networks on Facebook of university students and found that sharing the same university, race, and interests were the strongest predictors of friendship. Other studies analyzed group network structures and found that increasing centralization of group leaders improves group performance (Levi et al. 1954) and that teams that occupy a central position within an inter-group network perform better (e.g., Raz and Gloor 2007, Cross and Cummings 2003, and Balkundi and Harrison 2006). The hypotheses in Nann et al. (2009) were motivated primarily by these studies.

4.2. Overview of the Literature, Clustered by Years

It is not difficult to see that the number of publications on "Investor Sentiment," "Predictive Analytics," and "Big Data" has been growing. In the last ten years in particular, one can observe a positive relationship between increased use of social media and the number of academic studies about the impact and consequences of UGC.

As already stated in the introduction, the general idea that the emotions and psychology of investors influence the development of stock prices predates the rise of Twitter and other social media channels. Daniel et al. (2002), for instance, argued before Twitter existed that investors are human beings and make major systematic errors. They stated further that psychological biases and emotion have a substantial effect on stock market prices. While their arguments came from a more theoretical perspective and were not based on large empirical data analyses, the past 15 years have seen many academic studies in which their arguments are confirmed through large-scale analyses of data from Twitter and other social media.

Antweiler and Frank (2004) kicked off a string of publications – perhaps even a completely new stream of research – that followed in the next decade. Brown and Cliff (2004) or Qiu and Welch (2004) were also early contributors with their studies of investor sentiment and the stock market.

Nofsinger (2005) made a major contribution to the field with his publication on social mood and financial economics. Brown and Cliff (2005) extended their research with another study about the relationship of investor sentiment and asset prices. Das et al. (2005) and Whitelaw et al. (2005) were also among researchers contributing to the field during this time.

Baker and Wurgler (2006) presented another study in the *Journal of Finance* about investor sentiment and the cross-section of stock returns. Different papers anchored in behavioral finance were contributed by Bono and Ilies (2006), Tseng (2006), and Barber et al. (2006). Das and Chen (2007) presented a

widely cited paper in 2007 in *Management Science*. Other studies began to use Twitter (Java, 2007), and a growing number of papers looked more closely at investor sentiment, such as Tetlock (2007), Ku and Chen (2007), Baker and Wurgler (2007), Leskovec et al. (2007), and Mitchell and Mulherin (2007).

In 2008, the number of publications increased markedly. De Choudhury et al. (2008) and Sabherwal (2008), for example, looked at the relationship of online talk and stock market activity. Despite that Twitter had been founded two years earlier, most empirical studies focused on blogs or message boards (e.g., Pang and Lee 2008, Krauss et al. 2008). That year also saw many papers published on sentiment analysis and based on online communication (Abbasi et al. 2008, Tetlock et al. 2008, Barber et al. 2008).

The year 2009 revealed clearly increased interest in sentiment analysis and more sophisticated text analysis approaches to extract content from UGC (e.g., Melville et al. 2009, Boiy and Moens 2009, Choi et al. 2009, Lin and He 2009, Narayanan et al. 2009, Schumaker and Chen 2009). New media sources such as Twitter and Stocktwits became more popular for researchers such as Zeledon (2009). However, large empirical studies were still rare.

Researchers that same year began to look into this topic more seriously and worked with the new datasets. Prominent papers by Asur and Huberman (2010) about the predictive value of social media, a working paper by Sprenger and Welpe (2010) that resulted in a full paper in 2014, and a study by Bollen et al. (2010) of Twitter sentiment and socioeconomic phenomena were published. Then, in 2010, the number of papers that used Twitter and tweets as the central research element increased significantly; examples include Sriram et al. (2010), Bifet and Frank (2010), Ye and Wu (2010), Zhang et al. (2010), and Yang and Count (2010). Boyd et al. (2010) collected 720,000 tweets to study re-tweeting behavior on Twitter. There was also research into the number of followers of Twitter users and their influence (e.g., Cha et al. 2010, Kwak et al. 2010). Other papers focused on different sources, such as news and blogs (e.g., Gilbert and Karahalios 2010, Zhang and Swanson 2010). Deeper research into concrete trading models based on UGC was presented by, for example, Zhang and Skiena (2010) and Schumaker and Chen (2010).

In 2011, Bollen et al. (2011) presented their prominent and widely cited paper on how Twitter mood predicts the stock market. They collected approximately 9.85 million tweets and found predictive value for the DJIA. This was the largest Twitter dataset that had been used to date, and the work reverberated widely not only in academia but also in the finance industry itself. Another work by Bollen and Mao (2011) utilized Twitter to draw conclusions on stock predictions. Mao et al. (2011) compared the value for financial prediction of a range of online datasets (Twitter feeds, news headlines, and volumes of Google search queries). Oh and Sheng (2011) presented influential work at the ICIS conference on the predictive power of stock microblog sentiment. Notably, the focus in 2011 shifted slightly from empirical and more descriptive studies to papers that looked deeper into the specifics of the data and that applied more sophisticated text analysis approaches. For instance, Dion et al. (2011), Wang et al.

(2011), Chen et al. (2011), and Groß-Klußmann and Hautsch (2011) all presented new text mining approaches to understand better the sentiment in online messages. Others showed that text-based communication can spread emotions among group members (Guillory et al. 2011). Many authors were interested in a better understanding of the causality of new media on the stock market (Engelberg and Parsons 2011) and trading models based on the new datasets (e.g., Wang et al. 2011, Kara et al. 2011, Chan and Franklin 2011, Groth and Muntermann 2011). Others focused more on understanding the impact of influencers or experts on the quality of predictions (e.g., Bakshy et al. 2011, Hill and Ready-Campbell 2011).

The popularity of research into the effects of Twitter on the stock market continued in 2012, with papers by, for example, Brown (2012), Xu et al. (2012), Giannini et al. (2012), Valerius (2012), Ruiz et al. (2012), Tirunillai et al. (2012), and Vu et al. (2012). Interest in employing more in-depth sentiment analysis methods to explain the value of online talk on the stock market also remained high, as reflected in papers by, for example, Krauss et al. (2012), Brown (2012), Fang et al. (2012), Mizumoto et al. (2012), Schumaker et al. (2012), Rao and Srivastava (2012), and Vu et al. (2012). Kramer 2012 took a slightly different spin and researched the spread of emotions via Facebook. Other researchers investigated the Twitter network structure in more detail. Lerman et al. (2012) stated that followers, friends, and re-tweets lead to a large social network in which news stories and other content can easily spread.

Researchers began in 2013 to explore more specific hypotheses. Contribution IV Nann et al. (2013), for instance, used longer time series than in previous research to build a trading strategy on intraday sentiment indicators. Prior studies had looked only at daily values. Papaioannou et al. (2013) also used intraday prices to build a forecasting model for exchange rates. Hagenau et al. (2013) and Makrehchi et al. (2013) applied advanced sentiment methods, such as using context-capturing features and event-based approaches, to extend prior research. That same year, seven years after Twitter was founded, some researchers published critical studies about the quality of Twitter data and its users (D'Onfro 2013, Lee 2013).[6] Alternative sources became popular to find correlations with the stock market. Xu and Zhang (2013), for example, looked into the impact of Wikipedia on market information environment. Moat et al. (2013) analyzed Wikipedia to find usage patterns before stock market moves Preis et al. (2013) used Google Trends to quantify trading behavior in financial markets. Gagnon (2013), Siering (2013), Li and Li (2013), Alanyali et al. (2013), and Oliveira et al. (2013) also contributed to the research that year.

The number of papers in the field of predictive analytics based on social media and UGC hit a new high in 2014. There were many studies about Twitter and stocks, such as Rao and Srivastava (2014), Sprenger et al. (2014), Chen et al. (2014), Nofer and Hinz (2014), Li et al. (2014), Si et al. (2014), Ding et al.

[6] http://www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10 or AP Twitter account hacked in market moving attack. Bloomberg. https://www.bloomberg.com/news/articles/2013-04-23/dow-jones-drops-recovers-after-false-report-on-ap-twitter-page. Accessed 24 April 2013.

(2014), Li et al. (2014), Curme et al. (2014), Kim and Kim (2014), Ammann et al. (2014), Q. Li et al. (2014), de Fortuny et al. (2014), and Geva and Zahavi (2014). There were also studies about the predictability of exchange rates, including Crone and Koeppel (2014), and Janetzko (2014). Christakis and Fowler (2014) and Kramer et al. (2014) presented comprehensive studies about the detection of emotional contagion in social networks. Other authors focused on text mining and sentiment analysis to find evidence of the predictive value of news or messages posted on social networks; these include, for example, Luo et al. (2014), Nassirtoussi et al. (2014), Zheludev et al. (2014), Q. Li et al. (2014), Jiang et al. (2014), Ammann et al. (2014), X. Li et al. (2014), Kearney and Liu (2014), Z. Li et al. (2014), Yang et al. (2014) and Gottschlich and Hinz (2014).

Measured by the number of publications, the research trend continued in 2015. Ranco et al. (2015) explored the effects of Twitter sentiment on the stock market. Thakkar and Patel (2015), noting that number of papers dealing with sentiment analysis was already fairly high in 2015 and remained quite popular among researchers, presented a summary of different sentiment analysis approaches on Twitter. At the same time, it became apparent that a growing number of Asian researchers were working on this research topic – a trend already hinted at in 2014. In 2015, there were papers from Zhou et al. (2015), Yuan (2015), Nguyen et al. (2015), Ding et al. (2015), and Liu et al. (2015), for example; they contributed significantly to the research topic. Other studies based on innovative extensions of earlier research dealt with the predictive value of Twitter and news on the stock market. For instance, Nofer and Hinz (2015) considered social interactions of Internet users to show the relationship between mood level and stock returns, which had not been done in detail before. Lillo et al. (2015) looked at a very specific subset of data to investigate the role of news in the trading and investment decisions of individual investors – a difficult area of research given the availability of microdata about the activity of individual investors. Liu et al. (2015) demonstrated that firms with official Twitter accounts have much higher comovement than firms without such accounts. They proposed a novel model to identify homogenous stock groups and predict comovement with respect to firm-specific microblogging metrics such as the firm's number of followers and the number of tweets sent.

There were further developments in using UGC or data from Twitter to explain movements on the stock market in 2016. Papers became more specialized as researchers focused on applying previously unused methods or using new sets of subdata (e.g., Q. Li et al. 2016, Dimpfl and Jank 2016, Shynkevich 2016) or included specific techniques or summaries (e.g., Kharde and Sonawane 2016, Kumar and Ravi 2016).

Significant research by McLean and Pontiff (2016), published in the *Journal of Finance*, raised the question: "Does academic research destroy stock return predictability?" Considering the number of publications on the topic over the previous decade and the relatively high number of different methods that had been applied to sentiment analysis and text mining approaches, their question was definitely of high interest to researchers. McLean and Pontiff (2016) identified 97 characteristics and used them to explain cross-sectional stock returns in peer-reviewed finance, accounting, and economics journals.

They found portfolio returns were 26 percent lower out-of-sample and 58 percent lower post-publication. The out-of-sample decline was an upper-bound estimate of data mining effects. Their findings suggested that investors learned about mispricing from academic publications.

In 2017, one of the contributions of this dissertation was published in the proceedings of the ICIS conference in Seoul (Janetzko et al. 2017). With a focus on a predictive model based on different sources for EUR/USD trading, the paper made contributions unlike any in previous research. Other papers from 2017 were similarly to papers from 2016, focusing on specific features, sub-datasets of techniques, and methods that had not been applied the same way before. For example, Song et al. (2017) used learning-to-rank algorithms, Yang et al. (2017) used genetic programming to optimize for a sentiment feedback strength based trading strategy, and Deng et a. (2017) adapted sentiment lexicons to domain-specific social media texts.

Based on the literature review, it can be seen that the number of papers decreased after 2015. The general direction of papers in 2018 aimed to provide a broader and more in-depth understanding of investor sentiment (e.g., Mahmoudi et al. 2018). Feuerriegel and Gordon (2018) focused on longer-term forecasting models based on regulatory disclosures. The same authors also presented a paper on news-based forecasts of macroeconomic indicators (Feuerriegel and Gordon 2019). Li et al (2018) reviewed systematically some 229 research articles on quantifying the interplay between Web media and stock markets from the fields of finance, management information systems, and computer science. Their stated goal was to clarify then current cutting-edge research to understand fully the mechanisms of Web information percolation and its impact on stock markets from the perspectives of investors' cognitive behaviors, corporate governance, and stock market regulation. Other papers published in 2018 may not have been available at the time of this literature review.

The next section provides a brief overview of the company StockPulse and its relationship to this dissertation. It is followed by a historical comparison of the papers presented here, in chronological order, and the aggregated number of news and messages from social media collected by StockPulse over the course of the last decade. The aim is to show the positive relationship between the increased usage of social media and the number of publications in the field of sentiment analysis and predictive analytics focusing on data from social media.

In addition, there is a detailed description and more in-depth explanation of the data, some of which was precluded from the contributions presented in this dissertation primarily because of page and format restrictions. Nevertheless, it is important to shed greater light on this to obtain a better understanding of the basic functionality of the data and their value for further use and application. How the data are analyzed is also a major underlying component of some of the papers (Janetzko et al. 2017; Nann et al. 2013).

5. StockPulse and the Relationship to this Dissertation

In a 2007 Master's seminar on "Collaborative Innovation Networks" (COINs) at the University of Cologne, Jonas Krauß and Stefan Nann, the author of this dissertation, were both assigned to the same working group. The goal was to find an online source that provided public data that could be used to derive predictive insights, and the two men settled on the Internet Movie Database (IMDb, https://www.imdb.com/) and, more specifically, IMDB's community message boards. On these message boards, IMDB users discuss movies, actors, potential Academy Award winners, and many other topics related to movies and Hollywood.

The seminar work focused on a particular sub-message board, "Oscar Buzz," where people were talking very specifically about the chances of given movies and actors winning an Oscar at the upcoming Academy Awards. The project involved collecting more than a year's worth of messages, applying text analysis methods to those messages to identify mentions of movies, and determining a numerical value for every movie based on that sentiment analysis. This resulted in an aggregated value for every movie, which could then be compared with others to create a ranking. Nine of the top-ten ranked movies actually won an Oscar at the next awards ceremony. Eventually, the seminar work was extended to a full paper that was published in the proceedings of the European Conference on Information Systems 2008 held in Ireland (Krauss et al. 2008). A more detailed introduction to this work appears in section 7.

Based on this success, the research scope extended to other areas. Financial markets seemed to be a promising field in which to confirm the insights from the "Oscar paper" that remained a bit blurry. Krauss and the author of this dissertation have worked closely on the topic ever since, finding success along the way both academically and professionally. For instance, the present author's diploma thesis in 2008, "Prediction of Stock Prices by Analyzing Digital Social Networks," looked specifically at stock trader message boards on Yahoo! Finance (https://finance.yahoo.com/). The work was done at a time when Twitter and other stock microblogging channels had not been widely adopted by users or for academic research. The key results and findings of this diploma thesis encouraged both men to follow this path and further develop methodologies, software programs, and other applications.

After some more research in related areas, including at the Center for Collective Intelligence at Massachusetts Institute of Technology (MIT) in Cambridge – research that did not focus on financial topics (e.g,. Gloor et al. 2009, MIT working paper by Nann et al. 2010), the present author founded the company StockPulse (https://www.stockpulse.de/) to concentrate on the key idea for financial markets.

StockPulse specializes in Emotional Data Intelligence. The core idea is to improve and support data-driven decision making for institutional financial investors by collecting and analyzing alternative data from social media, with a particular focus on financial markets. While raw and unprocessed data provide only isolated and limited pieces of information, emotional data match and connect the dots and – most importantly – make it possible to measure emotion that is inherent to the information. Web crawlers are

continuously scanning thousands of different Internet sources for relevant financial topics and communication and collect several million tweets, chat messages, message board posts, news articles, and comments to news articles each day. Unstructured text documents are processed with methodologies from Natural Language Processing (NLP) and more advanced sentiment analysis methods to extract topics, for instance, with Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model which has been used to extract abstract topics from text documents. It investigates the hidden semantic structure from large amounts of text (Blei et al. 2003). NLP is done in German, English, and Chinese.

StockPulse clients come from different geographic and business areas and include top-tier hedge funds, asset managers, banks, stock exchanges, private equity companies, financial publishers, online brokers, news portals, research companies, and universities. They use the data for different purposes; for example, hedge funds are seeking trading signals or trading models that will allow them to extend their existing portfolios. Stock exchanges use the data for trading surveillance or market watch. Private equity companies want to find new investment opportunities based on alternative data that are not yet accessible to everyone. The data can be accessed in various ways to improve decision-making processes, such as via an application programming interface (API) or a Web-based dashboard.

Two of the individual contributions of this dissertation (Nann et al. 2013 and Janetzko et al. 2017) used large-scale datasets and analyses from StockPulse. Section 7 describes these studies and the datasets used in more detail.

6. Data Specifications and Analytics

6.1. Comparison of Published Research and Social Media Activity

Figure 3 illustrates the number of publications on sentiment analysis and investor sentiment as described earlier in the literature review. While it is in no way meant to imply a complete picture, as it is possible some relevant papers were missed in the literature search, the chart does show the increasing interest of academics over time. What began with only a handful of publications in 2004 and 2005 increased steadily over the following ten years. It clearly peaked in 2014, a year in which close to 40 publications were identified. It is surely the case that academic conferences and journals adjusted their tracks and workshops to account for the growing interest for the topic.

The strong decline after 2014 shows a saturation of the topic. In the last few years, there have been fewer publications, but their focus has been more specific. Note that papers published beginning in 2018 may not be represented because they were not yet available when this analysis was completed.

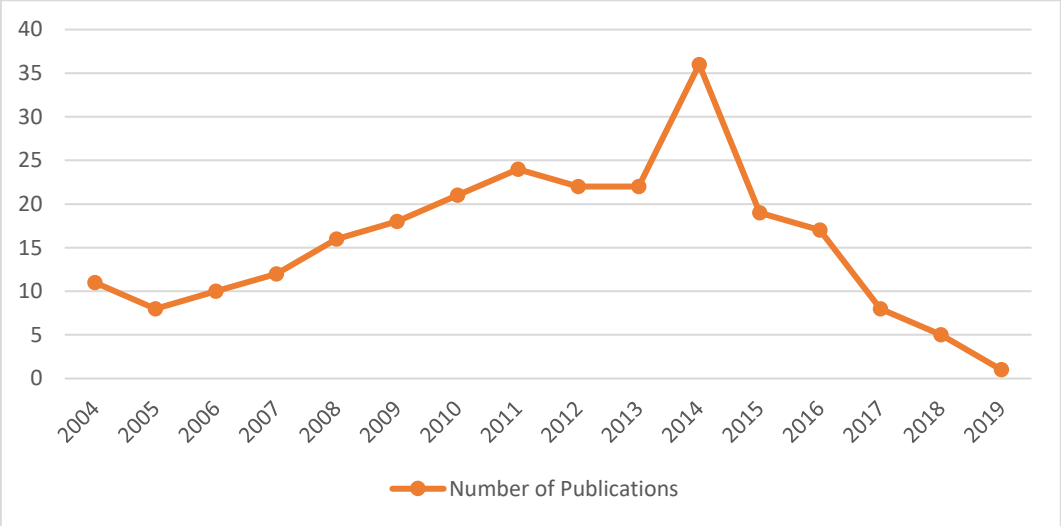

Figure 3: Number of publications based on the literature review

Figure 4 shows a historical distribution of messages collected by the StockPulse crawlers. Since the beginning of 2011, the crawlers have run nearly continuously, without major interruptions. Therefore, it is possible to get a good overview of the increased activity of financially relevant discussions in social media over that period.

While the crawlers found only 10,000 messages per day in early 2011, today there are more than 1.2 million messages each day. StockPulse is always improving the quality and capacity of these programs, which means that larger spikes may be due to new or improved crawlers. In addition, new data sources are regularly discovered.

Although, one cannot directly compare the time scales of the two charts with each other (the time scale of figure 4 is shorter and more granular than that of figure 3), we can see that there is a positive

relationship between the increasing number of publications from 2011 until 2014 and the increase in messages during the same time.

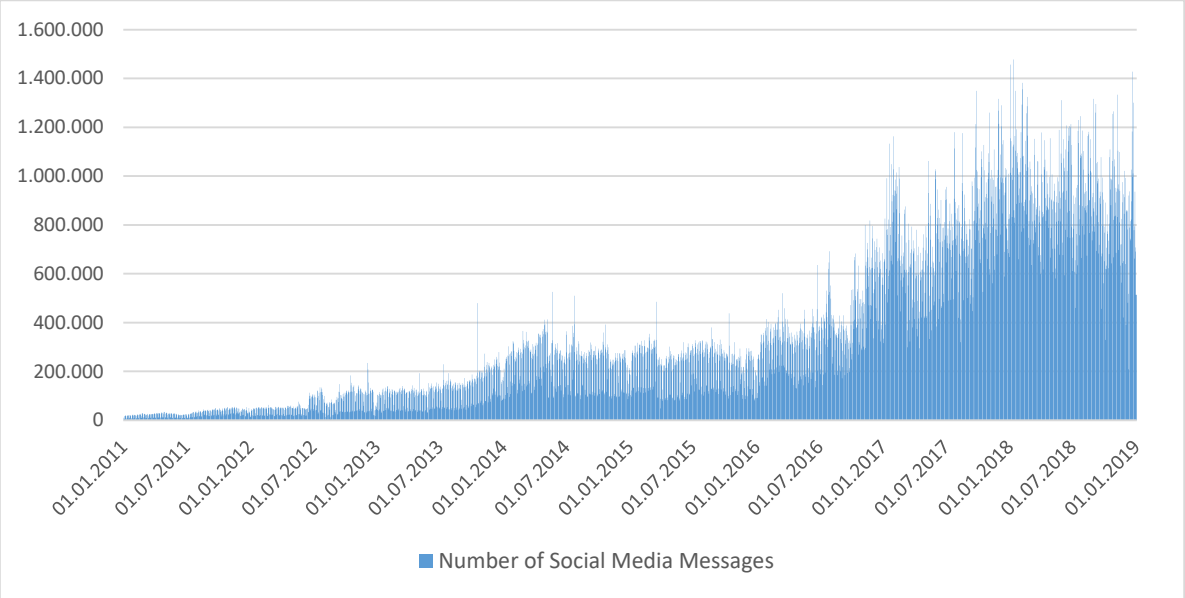

Figure 4: Number of messages collected by StockPulse since 2011

6.2. Data Description and Specifications

Page restrictions in the published papers precluded detailed specifications of the data and more thorough explanations of the sentiment analysis methods applied. This section compensates for those restrictions and provides a more detailed description of all main processing steps and special adaptations of the data to build the backtesting model in Janetzko et al. (2017) and the trading model in Nann et al. (2013).

Crawling many different unstructured sources on the Web requires sophisticated software routines – StockPulse has built and optimized processes over the last ten years. The StockPulse systems monitor worldwide online communities for more than 40,000 financial instruments, stocks, major indices, foreign exchange rates, commodities, or crypto currencies. The systems include complex matching and filtering processes to assign a text document (e.g., a tweet or message board post) to a financial instrument. Before the final sentiment values are calculated, another step aggregates filtered and classified text documents in different time frames. The following paragraphs provide short descriptions of the different steps.

6.3. Web Crawling

Web crawlers are programs designed specifically to access online sources and collect text documents (StockPulse's crawlers access only publicly available sources and data). There are some technical differences in how data from different sources is accessed and obtained. Twitter, for instance, can be accessed via a publicly available streaming API, whereas other sources, such as specific trading message

boards, must be accessed with individually adapted scraping programs that extract relevant information (e.g., name of the user or author, timestamp, and article content).

6.4. Filtering

Sophisticated matching algorithms detect ticker symbols, stocks IDs such as ISIN codes, Bloomberg and Reuters tickers, company names, and derivatives of company names. This method assigns every document to one or more entities, which is an important step because the quality of the calculations that follow depends on messages being correctly assigned. The system also maintains comprehensive identifier-mapping tables for financial instruments and indices worldwide. Identifier-mapping is the assignment of relevant identifiers to a document based on recognized entities; for example, a news document discussing an analyst's upgrade of automobile manufacturer Tesla would be tagged with the Bloomberg ticker TSLA:US.

In a next step, a spam filter algorithm scans all messages for insulting words or phrases (so-called "rants" or "flames"). Most posts of this kind can be identified by their scurrilous or nasty language. There is also a more advanced approach that analyzes relationships between users (e.g., in the case of Twitter, analyzing the follower network of users or monitoring interactions such as likes, mentions, or re-tweets) and calculates a reputation rank or "author score" for every user. The impact of every message is based on this internal author score. There are manually curated and verified social media users (financial experts or renowned news agencies) who carry a higher author score by default. Elon Musk, Warren Buffet, and the Bloomberg news agency, for instance, are assigned to this category because their tweets typically have a higher impact compared to the posts of lesser-known users.

Bag-of-words and Naive Bayes approaches are used in the natural language processing task to identify positive and negative connotations of a document. Both positive and negative bag-of-words are curated and maintained manually by people with a deep understanding of financial market-related communication. Both bag-of-words consist of several hundreds words each. Text analysis has been developed specifically for communication in financial markets, including for the informal language that is often used in UGC. For example, the words "long" and "short" have specific meanings in financial markets: "long" refers to buying a stock and "short" refers to selling. Because both words have different meanings in other contexts, the NLP task reduces noise and mismatches by focusing on context-specific analysis (Krauss et al. 2012). It is optimized to understand better the mood reflected in social media about stock market-listed companies. The NLP task also automatically identifies different languages; at present, it understands German, Chinese (traditional and simplified Mandarin), and English. This sub-task is very important since it translates words into numbers, which are again used in subsequent calculation processes.

Another task in the NLP module is Latent Dirichlet Allocation (LDA). Generally, LDA extracts topics from a given text document and ranks them by different importance weights. LDA posits that each

document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

6.5. Aggregation

All messages obtained through previous processing steps are summarized on different time scales. The term "report" is used in the following subsections to describe the aggregation of multiple message statistics for a certain time frame. Reports can be classified into three main categories that define their aggregation level and weighting.

6.5.1. Unweighted Reports

Unweighted reports summarize the number of posted messages:

- Raw Volume Reports: These are messages collected and assigned to a specific title. There are three different types of raw volume reports: (1) "Raw Total" define the absolute and total number of messages per title without sentiment; (2) "Raw Positive" define all messages identified with a positive sentiment score; and (3) "Raw Negative" define all messages identified with a negative sentiment score.

6.5.2. Exponentially Weighted Reports

Exponentially weighted reports use exponential smoothing to model the fact that a message has a certain "half-life" (explained in 6.6, below).

- Smoothed Volume Reports: These are exponentially smoothed messages based on raw volume reports. As in the previous category, there are three different types of these reports: (1) "Smoothed Total" define the smoothed total number of total messages; (2) "Smoothed Positive" define the smoothed positive message aggregations; and (3) "Smooth Negative" define the smoothed negative message aggregations.
- Median Smoothed Volume Reports: These are exponentially smoothed messages based on the median of smoothed values: (1) "Average Smoothed Total"; (2) "Average Smoothed Positive"; and (3) "Average Smoothed Negative.

6.5.3. Normalized and Weighted Reports

Normalized and weighted reports use exponentially weighted reports to determine relative measures.

- Weighted Volume Reports (Buzz): Buzz measures communication intensity, a weighted and relative measure for the current volume of messages compared to the median of smoothed volume. Based on Buzz, different titles are comparable with each other, for example, as a proxy for current market attention.
- Weighted Tonality Reports (Sentiment): Sentiment is a weighted and relative measure for positive and negative message counts compared to their median smoothed volume.

Not all reports are necessarily equal with respect to how much they aid in building prediction models or determining valuable signals. In some contributions of this dissertation, especially Nann et al. (2013) and Janetzko et al. (2017), different reports and their predictive values were used for data analysis. Figure 5 summarizes the process of crawling, filtering, and aggregating documents from social media and news sources.

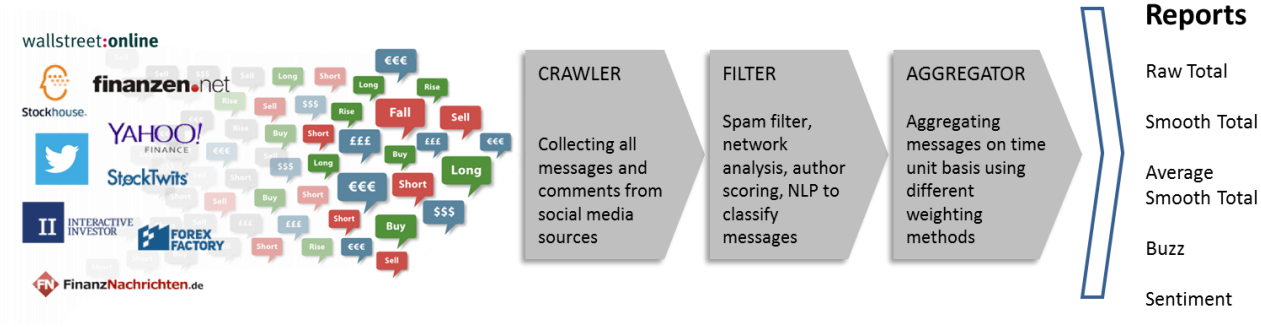

Figure 5: Crawling – Filtering – Aggregating messages from social media

6.6. "Half-life" of a Message

Figure 6 compares Raw Positive and Raw Negative reports from the beginning of the data coverage in 2011 until the end of 2018. It uses an average for all index constituents of the DJIA. One can observe that the number of average positive messages tends to be higher, on average, than its counterpart of aggregated negative messages. The same behavior can be observed for other titles and reports. It is important to be aware of this information when working with the data. Depending on the use case of the data, it may be necessary to apply measures to adjust the negative messages (e.g., to use a factor to weigh negative messages higher).

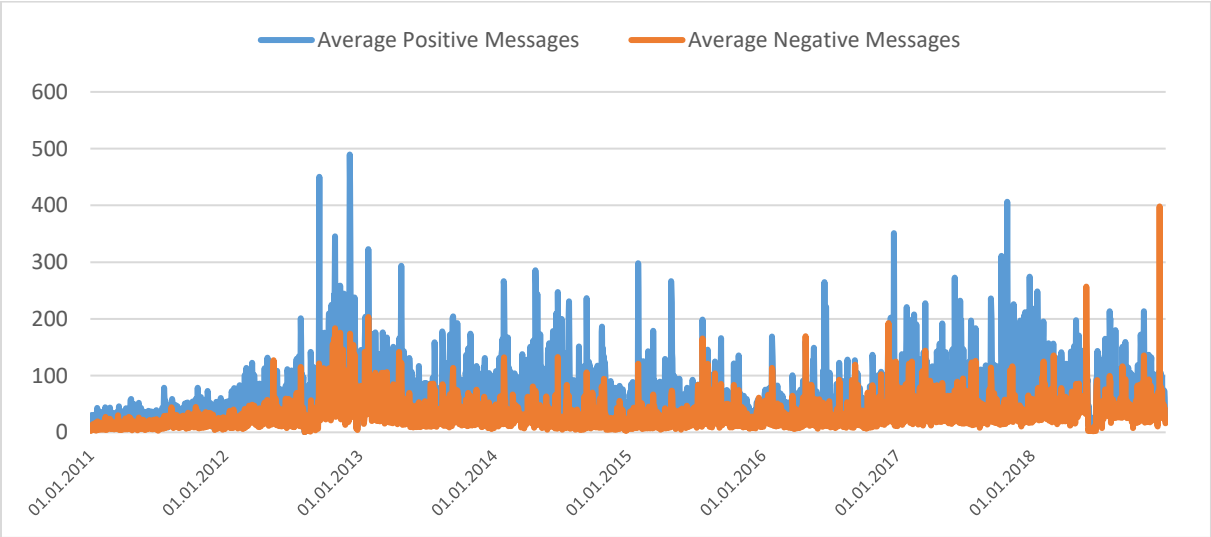

Figure 6: Average positive and negative messages for components of the DJIA

The number of users who only read social media messages is larger by far than the number of actively contributing users who write messages and tweets. However, it may be the case that all passive users

make decisions based on the comments they read. Once a message is published, it takes some time before it comes to the attention of the majority of interested (active or passive) users. After reaching an attention peak, the chance that a given message will draw more attention from additional readers declines. Messages move down on lists of "recent" posts over even a short time and eventually they become largely unseen by readers. Thus, a message has a lifecycle of potential to influence the decision making process. This phenomenon can be referred to as the "half-life" of a message.

Typically, the peak of attention is reached quickly while its decay takes longer. Usually, the older the message, the less attention it will receive and the less its potential to influence decision making. Researchers have explored this phenomenon. For example, Chen et al. (2003) proposed an aging theory to model the lifespan of news. They found that a news event becomes popular with a burst of news reports and fades away with time. Lu et al. (2019) did similar work on the lifespan of news events on Twitter.

The half-life of the potential for influencing decision making is reflected in the sentiment indicators. Exponential smoothing is a common measure to represent this fact mathematically with a formula. One important part of exponential smoothing is the "alpha"-factor (not to be confused with ALPHA information in financial markets), which is used to decay older input exponentially. Alpha, the case of financial markets, can be understood as a measure to weigh individual events in time according to their age or weigh a message's potential to influence decisions depending on when it was published. Low alpha means that the potential of more recently published messages is weighted lower; high alpha means that the potential of newer messages receives higher weight. In other words, high alpha means that more recent messages in social media are more relevant. While calculating the reports in the aggregation phase, the exact timestamp of every message is considered and the time difference of a given message to the calculation time of the respective report is used to decay its potential over time.

Figure 7 shows hourly sentiment data over time and compares high (0.66) and low (0.0561) alpha values. It aggregates the smoothed number of total messages for the German benchmark index DAX from January 2016 through December 2018 on an hourly basis ("Smoothed Total Reports"). Reports with higher alpha values tend to be higher than reports with lower values, which is depicted in the chart with consistently higher blue bars (blue time series = alpha of 0.66). For lower alpha values, messages from the past are weighted higher, which means that aggregated reports are smoothed stronger, resulting in lower values on average.

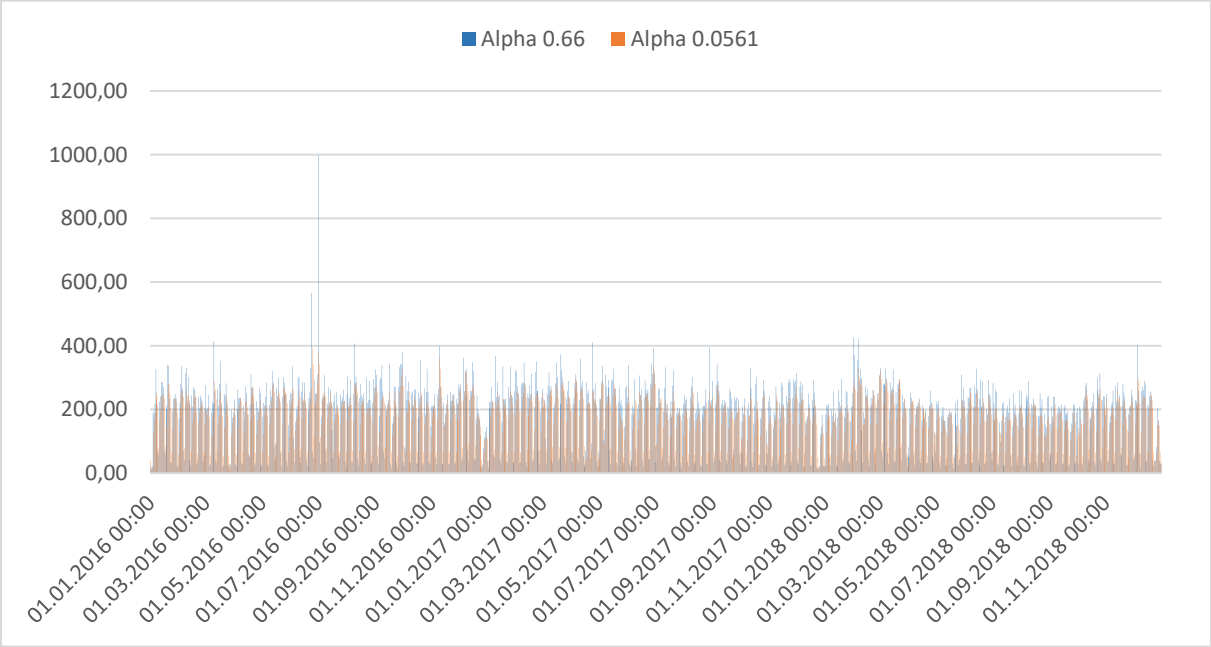

Figure 7: Comparison of hourly data with different alpha values for DAX Index

Spam filtering, context-specific NLP, aggregation, exponential smoothing, and consideration of half-life of text documents are important factors for the high quality of the sentiment indicators that were major components in different contributions of this dissertation: Contribution I (Krauss et al. 2008), Contribution IV (Nann et al. 2013), and Contribution V (Janetzko et al. 2017).

7. Summary of Individual Contributions

7.1. Overview

This section starts with an overview of the individual contributions of this dissertation, see table 1 for the complete list. For every contribution it provides the list of co-authors, the title of the work, and the publication type. The articles are sorted in ascending chronological order of their publication date. All articles were created in collaboration with co-authors. In section 7.2 there is a more detailed description of the work with co-authors and the contribution of the author of this dissertation for every paper. Section 7.3 presents an overview of the research topics and how the individual articles are embedded in this context. In the following section 7.4 there is a presentation of every contribution. It starts with a summary of every article, its requirements, results, and conclusion. Furthermore, there is a description of the position and contribution of every article to the overall research context.

Contribution	Authors	Title (Year)	Publication	Type
I	Jonas Krauß Stefan Nann Daniel Simon Kai Fischbach Peter A. Gloor	Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis (2008)	European Conference on Information Systems (ECIS)	Completed Research Paper
II	Peter A. Gloor Jonas Krauß Stefan Nann Kai Fischbach Detlef Schoder	Web Science 2.0: Identifying trends through semantic social network analysis (2009)	International Conference on Computational Science and Engineering (IEEE)	Completed Research Paper
III	Stefan Nann Jonas Krauß Michael Schober Peter A. Gloor Kai Fischbach Hauke Führes	Comparing the structure of virtual entrepreneur networks with business effectiveness (2009) and The Power of Alumni Networks - Success of Startup Companies Correlates With Online Social Network Structure of Its Founders (2010)	Collaborative Innovation Networks (COINs) and MIT Sloan School of Management; MIT Center for Collective Intelligence	Completed Research Paper and Working Paper
IV	Stefan Nann Jonas Krauß Detlef Schoder	Predictive Power on Public Data – the Case of Stock Markets (2013)	European Conference on Information Systems (ECIS)	Completed Research Paper
V	Dietmar Janetzko Jonas Krauß Stefan Nann Detlef Schoder	Breakdown: Predictive Values of tweets, Forums and News in EUR/USD Trading (2017)	International Conference on Information Systems (ICIS)	Completed Research Paper

Table 1. Overview of individual contributions

7.2. Work with Co-Authors and Contribution of the Author of this Dissertation

This dissertation consists of five individual contributions. The author of this dissertation is Stefan Nann. Jonas Krauß and Stefan Nann, who also founded the company StockPulse (see section 5), worked jointly on all articles. The order of the authorship of Jonas Krauß and Stefan Nann are interchangeable since the work was evenly shared between both with regard to an individual contribution. Stefan Nann has been first author of two of five individual contributions. **Contribution I** (Krauss et al. 2008) has five authors. Stefan Nann came up with the idea, literature research, data collection, construction of the methodology, and writing major parts of the article. **Contribution II** (Gloor et al. 2009) has also five authors. Stefan Nann came up with the idea for the second part of the article, the correlation of Web buzz with share prices. He also contributed to obtain results with network analysis software Condor, supported to organize literature research, and writing of the article, especially in the second part. **Contribution III** (Nann et al. 2009) has six authors. Stefan Nann had the idea, contributed to literature research, programmed Web crawlers to collect the data, developed data analysis methodologies, and wrote major parts of the article. **Contribution IV** (Nann et al. 2013) has three authors. The idea is from Stefan Nann and Jonas Krauß. The data was collected with Web crawlers created by Stefan Nann and Jonas Krauß. Stefan Nann contributed to literature research and both designed data analysis methodologies and came up with the trading model presented in the article. The text was collaboratively written by Stefan Nann and Jonas Krauß. **Contribution V** (Janetzko et al. 2017) has four authors. Stefan Nann contributed to the general idea, literature review, data collection, and writing of the article.

7.3. Context of the Research Theory

This section explains the research context and underlying research theories. The main research streams are set in relation to each other and the individual contributions are located in the relation matrix (figure 8). All contributions of this dissertation can be positioned in the theories of Social Network Analysis (SNA), sentiment analysis, and predictive analytics.

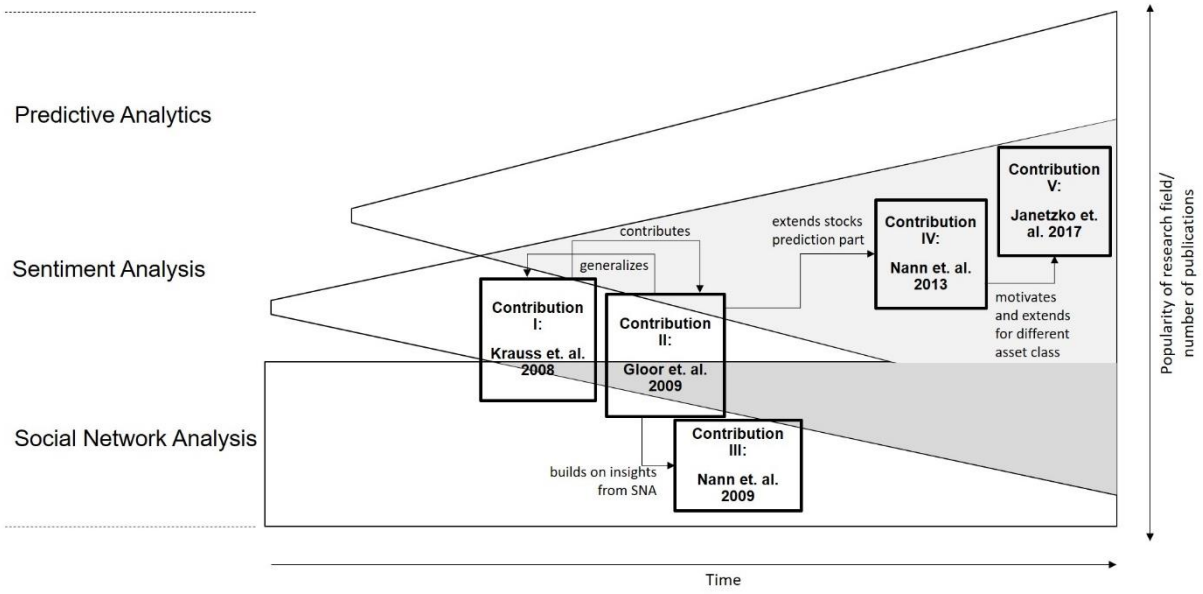

Figure 8: Main research streams and position of individual contributions

Social Network Analysis is the oldest research field of the research streams illustrated in figure 8. It has its theoretical roots in the work of early sociologists such as Georg Simmel (Simmel 1950) and Émile Durkheim (Durkheim 1960), who wrote about the importance of studying patterns of relationships that connect social actors. Bavelas (1950), Bavelas and Barrett (1951) and Leavitt (1951) introduce an important element of centrality in network analysis. Freeman (1979) and also Wasserman and Faust (1994) base their work essentially on this pioneering work and conduct a detailed discussion of the concept of betweenness centrality. Contribution I (Krauss et al. 2008), Contribution II (Gloor et al. 2009) and Contribution III (Nann et al. 20109) use the concept of betweenness centrality as a major element to identify trends based on SNA. Erikson (2013) states that the popularity of the idea of "networking," beginning in the late 1970s, and the rapid proliferation of social networking websites in the twenty-first century have magnified public awareness and interest in social network research.

By conducting a computer-assisted literature review Mäntylä et al. (2016) found out that initial research in sentiment analysis dates back to the beginning of 20th century. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web. Consequently, 99 percent of the papers have been published after 2004. This is also confirmed by the literature review of this dissertation and the number of articles presented in the clustered literature review (section 4.1). All individual contributions use methodologies of automated sentiment analysis to underpin their findings. Especially, Contribution IV (Nann et al. 2013) and Contribution V (Janetzko et al. 2017) use advanced methodologies of context-specifc sentiment analysis to build trading or backtesting models for financial predictions.

The field of predictive analytics is the youngest stream of research compared in this context. Although, its roots are dated back to the 1940s (SalesChoice 2016), the field moved into focus with the continuous

growth of stored data, combined with an increasing interest in using data to gain research and business related insights. Antweiler and Frank (2004) provided a key article about investor sentiment and a combination of theories of sentiment analysis and predictive analytics. The article has also been the basis for the stock market related articles of this dissertation. While Contribution I (Krauss et al. 2008) and Contribution II (Gloor et al. 2009) are using basic methodologies of applied predictive analytics, Contribution IV (Nann et al. 2013) and Contribution V (Janetzko et al. 2017) build more sophisticated trading and backtesting models to create insights about the relationship of emotional sentiment data and stock (Nann et al. 2013) or forex (Janetzko et al. 2017) market movements.

Figure 8 also illustrates the development of each research stream over time. While SNA remains at a steady popularity among researchers from social and computational sicences over time, the fields of sentiment analysis and predictive analytics receive an increasing attention over time. Particularly, over the last decade both fields have a prominent and growing standing in IS research.

The next section presents every article in more detail. It starts with the general idea of each article, its requirements, data collection approaches, data analysis methodologies, and short presentation of obtained insights. Eventually, the position and contribution of every article in the overall research context is elaborated in more detail.

7.4. Contribution I

Krauss, J.; Nann, S.; Simon, D.; Fischbach, K. and Gloor, P.A. (2008) *"Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis"* was published in the proceedings of the European Conference on Information Systems (ECIS) as a completed research paper.

7.4.1. Summary

Contribution I (Krauss et al. 2008) resulted from a course about Collaborative Innovation Networks (COINs) in the winter semester 2006/2007 held at the department of Information Systems and Information Management at University of Cologne. The goal of the course was to identify an online source which can be used to predict trends based on methods from SNA and sentiment analysis. The seminar work was extended with additional research after the course. Two online sources built the center of the analyses: the Internet Movie Database (IMDb) and "Box Office Mojo" (https://www.boxofficemojo.com/). Before IMDb closed its message boards on February 20, 2017 it was the largest community for movie and cinema related communication in the Internet. According to Big Boards the website had over 4 million users in 2007 and at least 15 million monthly unique U.S. visitors (Compete 2008). The webpage Box Office Mojo was used to extract fundamental data about analyzed movies, e.g., movie release dates and show times.

Contribution I follows two main hypotheses. Firstly, it assumes that the chances of a movie to win an Oscar can be determined by the communication structure of the IMDb community. The second

hypothesis assumes that there is a relationship between the communication intensity about a movie and the performance of the movie at the box office (Krauss et al. 2008).

The approach of the article was twofold. By mining text documents, i.e., message board posts from IMDb the discussion content was analyzed. A theoretical model based on three components was constructed: discussion intensity, positivity, and time. The content analysis was combined with information about the structure of the social network formed by users of the IMDb community. SNA tool Condor was used to analyze network structures of IMDb users. The software tool creates graphical static link views of the communication between users in a Web forum and calculates an actor contribution index, which delivers clues about the relevance and importance of key actors contributing to the communication (Gloor et al. 2003). Furthermore, Condor supports content analysis of terms being used in unstructured text documents.

Regarding the content analysis three indices were introduced: the "Intensity Index," the "Positivity Index," and a temporal noise factor, the "Time Noise Factor." The Intensity Index counts mentions of a specific movie in the text documents. The construction of this index followed the approach of Antweiler and Frank (2004) who found a significant correlation between the amount of messages being posted about stocks in finance-related online forums and their volatility. The Positivity Index used the software Condor for text analysis in order to identify whether the discussion about a movie was associated with positive terms. The terms have been determined by ranking potential phrases with regard to their betweenness centrality (Wasserman and Faust 1994) with Condor's content processing function. The third component is the Time Noise Factor. It considers the fact that some movies are released earlier in the year than others. Research has shown that it is a common industry practice to delay the release date of a movie until late in the year to improve its chances winning an Oscar (Nelson et al. 2001). The final "Oscar Model" is a combination of all three indices.

The results suggested that movies that are being discussed in a positive way in the message boards have a high probability of getting a nomination for the Academy Awards. Furthermore, results indicated that users who are participating in the communication are movie enthusiasts who value similar criteria in a movie as suggested in the Oscar poll.

Based on the positive results of the content analysis further research about the correlation between movie success and community communication structures has been conducted. The goal was to develop an appropriate metric to measure communication behaviour of the community regarding movies. In order to achieve this goal, a new index has been introduced: the "Trendsetter Index." This new metric was combined with the previously created Intensity and Positivity Index to form the final "Buzz Model."

The Buzz Model indicated that positive discussions about a movie in the message board have strong relationship with higher revenue of the movie at the box office. Robust support was found for the second hypothesis that higher movie success correlates with higher communication intensity (Krauss et al.

2008). It could also be observed that most influential users who had the highest betweennes centrality measures led discussions in the IMDb message board.

7.4.2. Contribution and Position in Research Context

Contribution I provides an extension of the research on the influence of online communities on the success of movies. It uses a combination of communication (sentiment) analysis and methodologies from SNA to build a model for the prediction of movie success. Figure 8 illustrates that Contribution I has a major overlap for the fields of sentiment analysis and SNA which has not been widely adopted by previous research in this field.

While IMDb has been frequently used as a basis to predict movie success by other researchers (e.g., Eliashberg and Sawhney 1996, Jensen and Neville 2002, Pardoe 2005, Simonoff and Sparrow 2000, Dellarocas et al. 2007, Kaplan 2006), little research has been done prior to the publication of Contribution I in using communication behavior and social networking structures of an online community as a determinant of movie success at the box office or as a predictor for Oscar nominations. For instance, Awad et al. (2004), analyze the influence of online movie ratings on box office success and developed statistical models based on these ratings to forecast movie revenues.

Contribution I extends prior research by considering social networking structures of an online community and by measuring discussion content with sentiment analysis rather than movie ratings. The measurement of the discussion content was done with text mining algorithms of software tool Condor (Gloor and Zhao 2004). The approach also involved a context-specific construction of sentiment words (bag-of-words) that describe the event of a movie winning an Oscar. This specific approach has also not been used in a similar way in this research domain before.

Findings of Contribution I have been the basis for parts of Contribution II which suggests a general model for the correlation of Web buzz and real world effects.

7.5. Contribution II

Gloor, P.A.; Krauss, J.; Nann, S.; Fischbach, K. and Schoder, D. (2009) *"Web Science 2.0: Identifying trends through semantic social network analysis"* was published in the proceedings of the International Conference on Computational Science and Engineering (IEEE) as a completed research paper.

7.5.1. Summary

Contribution II (Gloor et al. 2009) introduces a general model that maps buzz on the Web to real world events. The article proposes innovative methodologies, supported by Condor, to mine different sources on the Internet, e.g., news, blogs or forums and aims to identify trends and people launching these trends by applying algorithms from SNA and sentiment analysis.

The main goal was to built a "Web Buzz Index" for different topics that enables prediction of the popularity of the topics in the future. In the article the Web Buzz Index was applied to different concepts,

such as measuring the changes in popularity of brand names and famous people, e.g., movie stars, politicians, or business executives. Eventually, it was also used to spot correlations with stock prices.

The Web Buzz Index was constructed in a three-step process ("What – Who – How"). In the first step ("What") the betweenness centrality of the linking structure of a communication network was calculated. In the second step ("Who") the social network positions of the "actors" of a specific source were determined and added to the model that constructs the index. The third step ("How") applied text mining to the communication in order to understand its semantic context. The third step required the application of methodologies from sentiment analysis.

Analyses were supported by SNA software tool Condor (Gloor & Zhao 2004) which was originally developed to mine email networks to automatically generate dynamic social network movies. Condor includes algorithms to determine temporal network centrality measures, visualizing social networks as Cybermaps, and a semantic process of mining and analyzing large amounts of text documents (Gloor et al. 2003).

The first step ("What") measured temporal betweenness of concepts. Betweenness centrality in social network analysis tracks the number of geodesic paths through the entire network, which pass through the concept whose influence is measured (Gloor et al. 2009). As access to knowledge and information flow are means to gain and hold on to power, the betweenness centrality of a concept within its semantic network is a direct indicator of its influence (Wassermann and Faust 1994). Importance of concepts in a given context or information sphere can be determined by calculating their betweenness centrality. Betweenness centrality is usually represented as numerical values, ranging from zero to one. Zero indicates that there is no importance of the concept in the information sphere, a value of one indicates high importance. The article commenced with various examples of measuring betweenness centrality scores in a political context, i.e., the US presidential elections in 2006.

The "Who" step was based on the idea that certain people in a network or group possess a higher impact than other participants of the same network or group. The betweenness was used as an approximation for their impact. "Nodes" in a newtork are generally referred to as "actors." An actor can be a website, a blog, or a user in a message board.

The third step, coined „How" in the article, considered the semantic content of online-mediated communication. More specifically, it looked at positive and negative emotions in the discussion. Condor includes effective text analysis methods. Through its content process functionality the software automatically can identify most frequent words and word pairs in large amount of texts (Gloor et al. 2009). Different methodologies to extract the sentiment from text documents were applied in further steps, e.g., a bag-of-words approach, entity extraction and co-occurrence, or use of regular expressions. Sentiment analysis was conducted for finance-related documents which were collected from sub message boards that have been concerned with single stocks from Yahoo! Finance.

The Web Buzz Index is the combined measurement of the three steps. In order to test the predictability of the index it was applied to a finance-related use case. An algorithm that determined correlations between the index and actual stock price developments was implemented. Generally, final results have shown that a relation between Web buzz and stock price movements exist.

7.5.2. Contribution and Position in Research Context

The main goal of Contribution II was to construct a general model to show that buzz on the Web mirrors real world effects. Separating the Web into different sources and using metrics to determine the impact of actors on the Web permits to discover trends, in many cases before the real world has become aware of them (Kleinberg 2008). Figure 8 shows that Contribution II mainly builds on theories of SNA (Gloor and Zhao 2004). It also uses methods from sentiment analysis which have been adopted and developed further based on the findings of Contribution I. The second part of Contribution II about the correlation of Web buzz and stock prices also uses approaches of predictive analytics which is depicted with an overlap with this research field in figure 8.

The construction of the Oscar Index in Contribution I has been the basis of the Web Buzz Index that has been introduced in Contribution II to combine the general three step approach ("What – Who – How") to measure general Web buzz. It has been applied to stock market data in order to determine a correlation between Web buzz and stock price developments.

The type of building the Web Buzz Index has not been used in research before the publication of Contribution II. Especially, the adoption of different Web sources (websites, blogs, and forums) and the combination of network measures such as the concept of betweenness centrality with sentiment analysis approaches led to a unique model that revealed significant relationships between Web buzz and real-world developments.

7.6. Contribution III

Nann, S.; Krauss, J.; Schober, M.; Gloor, P.A.; Fischbach, K. and Führes, H. (2009) *"Comparing the structure of virtual entrepreneur networks with business effectiveness"* was published in the proceedings of the conference on Collaborative Innovation Networks (COINS) as a completed research paper.

Major work for this publication was done during a research program at MIT. The article is also officially listed as a MIT Sloan School of Management working paper under the title *"The Power of Alumni Networks - Success of Startup Companies Correlates With Online Social Network Structure of Its Founders"*.

7.6.1. Summary

Contribution III (Nann et al. 2009) studies entrepreneurial success of alumni of major German universities based on methods from SNA. The focus lies on a relationship network of entrepreneurs as it was represented on the German social networking site Xing (http://www.xing.com). The study

investigated if online social networking structures predict entrepreneurial success. Additionally, basic semantic matching algorithms were applied to classify user profiles and extract information about universities and professional occupation of entrepreneurs.

The article is based on four hypotheses. First two hypotheses focus on a network or group level and follow the assumption that the structure and position of a university alumni network have effects on the success of its entrepreneurial activity. Hypotheses three and four conduct observations on an individual level and are based on the assumption that the structure and position of individual entrepreneurs are important factors for their success.

The hypotheses that are based on a network level are motivated by research of Mayer and Puller (2008). The article expected to find cliques of alumni from the same university in the German founder network (old-boys networks). Mayer and Puller 2008 analyzed friendship networks on Facebook of university students and found that same university, race, and interests were the strongest predictors of friendship. Other studies analyzed group network structures and found that increasing centralization of group leaders improved the performance of the groups (Levi et al. 1954) or that teams which occupy a central position within the inter-group network performed better (e.g., Raz and Gloor 2007, Cross and Cummings 2004, and Balkundi and Harrison 2006).

On an individual level (hypotheses three and four) the article was mainly influenced by the study of Raz and Gloor (2007) who have shown that CEOs of startups are more successful if they communicate more with their peers. The article further assumed that entrepreneurs who are well embedded into the old-boys network of their university are more successful. Ahuja (2000) or Burt (2004) for instance have shown that people connecting structural holes are more successful. Murray (2004) suggested that academics who start biotech firms use their social capital to recruit collaborators through their local laboratory networks (Nann et al. 2009).

Data were collected with proprietary programmed Web crawlers from publicly accessible profiles from social networking platform Xing. Xing is the leading German business networking website, similar to LinkedIn. In total the crawlers collected 654,193 users and 4,456,393 relations between these users. By applying basic semantic matching algorithms on the collected user profiles they were classified by university. The matching included 12 German universities. In the next step all profiles were scanned for keywords that identified a user as an alumnus, e.g., Besitzer (owner), Unternehmer (entrepreneur), Gesellschafter (shareholder), Geschäftsführer (CEO), etc. This resulted in a sub-sample of the data of 15,143 user profiles and 232,390 relations.

Further statistics were collected to shed more light on the dataset. For instance the “graduating quotient” measured the number of students graduating per year compared to all students found in the dataset. The “founder quotient” was calculated as a percentage value of company founders and entrepreneurs compared to all alumni of a university. Both metrics were based on the collected Xing data.

For a manually selected sub-sample of 80 companies, that were identified in the previous processing step, more advanced measures were determined concerning their economic impact and company success. This helped to better understand the interrelationship between individual success and social networking behavior. Further network measures were calculated, such as group degree centrality and group betweenness centrality by using the software Condor. The article introduced a new concept for in-group networks referred to as “tribal networks” or “tribes”. The term “tribes” or “tribeness” was defined as the ratio of the number of actors and edges within the old-boys network to the number of actors and edges in the outside (external) network of a university (Nann et al. 2009).

The study subsequently analyzed different correlation values between tribal networks and the measures that were determined in the first part of the article, e.g., the founder quotient or the economic impact of entrepreneurs. Correlation values for full network metrics were also calculated. Full networks included, in addition to in-group relationships, also external relations of previously identified profiles with entrepreneurial characteristics.

Correlation indicated strong positive and significant results of tribeness with the relative economic impact of the university and the impact per founder. The findings also provided indications that it pays off to be a very closely connected community and creating a strong in-group “tribe” feeling and promoting strong bonding among alumni are means to success. Results on an individual level showed that metrics such as individual degree, weighted tribe factor, and betweenness centrality tribe factor are positively correlated with success. This means that a successful entrepreneur seemed to have proportionally more links with other alumni from her/his university than with external people.

The results of the article further suggested that university alumni networks that were successful in founding startups, measured by their average economic contribution, are organized as tribes. It could be found that their tribeness, the strength of their internal cohesiveness or their negative degree of openness to external actors correlated strongly with their economic success. For universities this could mean that they should motivate and encourage students to build up more and closer connections with their alumni (Nann et al. 2009). Findings on an individual level eventually showed that the more online friends an entrepreneur had, the more successful she or he was. Furthermore, the more tribal an entrepreneur was, the more successful she or he was with her or his company.

7.6.2. Contribution and Position in Research Context

Contribution III built on insights from network theory from Contribution II and confirmed that it is possible to make predictions based on publicly available social networking data. The application of centrality measures such as betweenness centrality and the use of software tool Condor were the basis for data analyses in Contribution III. As figure 8 illustrates, Contribution III is mainly positioned in the field of SNA theory as it used online social networking structures of entrepreneurs and universities to predict an entrepreneur’s or an entire network’s success. In order to classify the data and identify

company founders, for example based on their professional occupation, semantic analyses were applied. It also introduced a new concept called “tribes” or “tribeness” which defines the cohesiveness of a group or network.

While many authors have researched the effectiveness of groups or networks (e.g., Mayer and Puller 2008, Reagans and Zuckerman 2001, Balkundi and Harrisons 2006), Contribution III extended this research by using a completely new dataset which has not been used for a similar analysis before. The research setup of Mayer and Puller (2008) who studied social network structures of Facebook to analyze friendship networks of university students has been very similar to the methods used in Contribution III. However, the idea of using data from an online (social) business networking site to find indications for the performance of university alumni networks has been new and not used in research before.

The performance of individuals in networks was also subject of the study of Raz and Gloor (2007). They analyzed 100 software startups in Israel and found that the communication intensity of individuals (e.g. CEOs) with their peers significantly correlated with the probability of survival of the individual’s company. On an individual level, it was shown that people connecting structural holes are more successful (e.g., Ahuja 2000 or Burt 2004). Other researchers studied the success of individuals in their local networks (e.g., Murray 2004, Gulati 1995, McPherson et al. 2001, Simon and Warner 1992).

The findings of Contribution III on an individual level are unique because the dataset which was used to obtain the results has not been used before for these kind of studies. Findings have eventually shown that the more friends an entrepreneur accumulated in the online social network, the more successful she or he was. Furthermore, the more integrated into the local network an entrepreneur was, the more successful she or he was with her or his company.

7.7. Contribution IV

Nann, S.; Krauss, J. and Schoder, D. (2013) *“Predictive Power on Public Data – the Case of Stock Markets”* was published in the proceedings of European Conference on Information Systems (ECIS) as a completed research paper.

7.7.1. Summary

Contribution IV (Nann et al. 2013) collects and evaluates communication from multiple publicly available online sources, such as Twitter, online message boards, and traditional news pages. The aggregated data was processed with text mining and sentiment analysis methodologies to build a trading model based on a sentiment indicator. Daily close prices of the S&P 500 index were used as an independent variable to measure the predictive value of the indicator. Prediction quality was evaluated by applying a trading model that considered restrictions such as transaction costs or broker commission fees. The article provided evidence of predictive power of public data and built on theories from sentiment analysis and predictive analytics.

Many authors applied sentiment analysis to UGC from publicy available online sources to study the impact on the stock market. Antweiler and Frank (2004) set an early milestone in this research field with an article that was published in the *Journal of Finance*. They studied the predictive power of online message boards for the stock market by analyzing 1.5 million messages from Yahoo! Finance and Raging Bull.

Contribution IV added new insights to the research stream by providing a virtual trading model considering realistic conditions based on sentiment indicators which were extracted from public online communication. The primary goal of the article was to develop a working trading model based on a sentiment predictor condensed of public data. This was supported by the goal to achieve a positive return on investment.

The dataset consisted of nearly 3 Mio. messages related to stocks of the S&P 500 index. The data were collected from June 1, 2011 to November 30, 2011. It was one of the datasets with the largest history analyzed in this field compared to previous research. Proprietary matching algorithms were developed to assign messages to single stocks, e.g., by stock ticker symbol, “cashtags”, or name. After the matching step a spam filter was applied to clean the dataset. Proprietary programmed spam filters searched for messages which intended to insult other users without contributing any relevant information. Most of these messages could be identified by usage of bad language and were removed from the dataset. In the following text mining process different concepts were applied, e.g., Naïve Bayes classifiers with an adapted bag-of-words and part-of-speech tagging approach. One key finding of the article suggested that the quality of correct sentiment recognition depends on a context-specific application of the text mining methodologies. The more context-specific the algorithms were designed the higher the quality of sentiment recognition was (Krauss et al. 2012).

In order to create a common basis across all stocks from the considered universe, a simple moving average (SMA) for the sentiment values was applied. The article suggested that a simple moving average of 30 days (SMA30) is a good starting point. These values were used as a stock price predictor (“Sentiment Predictor”) which is defined as the sum of the ratios of current positive messages with the SMA30 of positive messages and current negative messages with the SMA30 of negative messages. By introducing the Sentiment Predictor it was possible to compare stocks that had different numbers of assigned messages and thus different levels of communication.

The results indicated that communication data collected from public online sources, which were analyzed with methodologies from sentiment analysis, have predictive power for future stock price developments. The findings confirmed results of previous research of Bollen et al. (2011), Sprenger and Welpe (2010), or Oh and Sheng (2011) for single source-based predictions and extended their validity to the case of using multiple source aggregations.

The main contribution of this contribution was the construction of a trading strategy that considered realistic environmental conditions such as transaction fees, commission costs or stock liquidity. Most previous research of the same field has not taken into account similar restrictions and discussed the findings in a more theoretical framework.

The trading model was based on several assumptions. Stocks had to be liquid and thus tradable, which was secured since the study was considering stocks from the largest US index, the S&P 500. Every stock was required to have an average trading volume of more than 3 Mio. shares per day and the stock price had to be higher than 10 US Dollars on the day the stock was selected for trading. For instance, a higher stock price normally ensures a lower spread which usually means lower trading costs.

For all stocks that met these criteria a daily overall return on investment (ROI) was calculated by summing up daily individual ROIs for these stocks. This was done for the entire period from June 1, 2011 to November 30, 2011. To simulate the model in a more realistic way, the ROIs were adjsuted with the general market movement in the same time period. The market was represented in this case by the SPY certificate which is one of the most liquid instruments on the financial market and accurately represents the current market value of the S&P 500 index.

The trading model registered 833 trades over the entire period and had a total ROI of 197,84 percent (adjusted by the market and without trading costs). Transaction costs were measured in basis points (bp). The results were reported for different levels of basis points (5 bp, 15 bp, 20 bp, 25 bp). Compared with the performance of the SPY certificate in the same time period (-5.22 percent) all results except one (25 bp) were better.

The result confirmed previous studies that investigated the relationship between online communication and financial markets and extended their validity to the case of multiple sources. Based on the findings of Contribution IV it was confirmed that predictive power rests in publicly available data – at least for data from financial markets.

7.7.2. Contribution and Position in Research Context

Contribution IV extended previous research to assess predictive power of social media talk by aggregating data from multiple sources (Twitter, several online message boards, and traditional news). It also used a comparable long historical dataset (six months) for analysis. Contribution IV used and extended insights of Contribution II in terms of using data from public online sources to predict future values on the stock market. It especially extended methods from sentiment analysis and built a more realistic trading model which revealed the predictive value on stock prices of financially relevant and publicly available communication data. Contribution IV is positioned in the intersection of theories of sentiment analysis and predictive analytics as figure 8 shows.

Other studies from this area as well as Contribution IV build on a combination of previous research from the fields of sentiment analysis and works related to predictive power of UGC. Sentiment analysis is a

very broad research field and has not only been applied to the financial domain. For a comprehensive review of literature see sections 4 and 7.3. Key articles for a joint usage of both fields were provided by Antweiler and Frank (2004) and later extended to a new set of data by Bollen et al. (2011). Contribution IV was mainly built on theories and methodologies used in these studies.

The dataset consisted of nearly 3 Mio. messages over a period of six months. This was one of the largest datasets analyzed in this field compared to previous research. Most research concerned with stock market predictions based on online data is mainly theoretical in nature and does not take into account real-world limitations such as broker fees, bid/ask differences, and liquidity. To demonstrate the potential practical application of the findings of Contribution IV, a trading model based on the predictor, considering commission fees, transaction costs, and stock liquidity was provided. Another key finding of Contribution IV suggested that the quality of correct sentiment recognition depends on a context-specific application of the text mining methodologies which was built on Krauss et al. 2012.

However, some limitations were identified and also addressed in Contribution IV. The time period which was used to construct the trading model was rather short. Certain market conditions that were present during the analyzed period were responsible for the positive ROI of the model (e.g., the period from July to September when signals performed significantly better than on average). Long-term studies (at least over multiple years) are required to cover different market and economic phases to show the robustness of the model.

Another shortcoming at this stage of the research was the comparison of models based on different data sources. Based on the assumption that Twitter, online message boards, and traditional news differ with respect to their predictive power for stock price movements Contribution V was created.

7.8. Contribution V

Janetzko, D.; Krauss, J.; Nann, S. and Schoder, D. (2017) *"Breakdown: Predictive Values of tweets, Forums and News in EUR/USD Trading"* was published in the proceedings of the International Conference on Information Systems (ICIS) as a completed research paper.

7.8.1. Summary

Contribution V (Janetzko et al. 2017) addressed a major shortcoming of Contribution IV (Nann et al. 2013) as well as those of other previous research. The comparability of the predictive value of different online data sources was a limitation of many studies in the same area. Contribution V therefore looked into the predictive value of different types of sources of public online data using a unifying framework to achieve comparability of results.

The focus was put on two research questions. Firstly, given a number of alternative data sources for predicting future exchange rates of EUR/USD which one facilitates the best prediction? Secondly, what are the more specific conditions that made the winning prediction perform best?

The data were collected from multiple popular social media sources and a backtesting approach was applied to each of them. This allowed for an objective comparison across different sources. The backtesting approach was designed to identify parameters that increase cumulative returns. The work also addressed robustness of the findings by using samples from a period of more than 40 months, which was considerably longer than those used in previous studies. Furthermore, samples consisted of intraday sentiment data allowing for a more realistic and more significant trading simulation. As the majority of previous research was concerned with stocks and indices, e.g., Tumarkin and Whitelaw (2001), Antweiler and Frank (2004), or Bollen et al. (2011), Contribution V adds value by studying sentiment-based predictions for exchange rates.

Data consisted of messages from Twittter, online forums and news sites and were collected in the period of June 19, 2012 and November 12, 2015 which was a long time period compared with other studies. The dataset was specifically comprised of messages with mentions of Euro, EUR, Dollar and USD. For instance on Twitter people use cashtags (represented as a dollar sign, $) as a prefix for ticker symbols such as $EURUSD to identify a financial instrument in a tweet. Other identifiers were used to match the Euro/US Dollar exchange rate in text documents. A full list of identifiers is provided in the appendix of the article. The next step of data processing identified and filtered spam. Additionally, several text mining methodologies were applied to determine the sentiment or tonality of each message towards the Euro/US Dollar exchange rate. Major features that were developed and applied in Contribution IV have been adopted and extended for Contribution V. In order to measure the quality of the sentiment score, precision recall ratios were calcuated. With a precision of 0.83 and a recall of 1 (every message was classified) the classifier achieved a quality measure F of ~0.907.

The article applied a backtesting approach to simulate estimated profits and losses, which an investment strategy could have earned in the past. The approach was used to generate buy and sell signals, in order to maximize returns of a simple cash portfolio by trading the exchange rate EUR/USD. A fixed-size moving window technique was applied to the sentiment values to generate trading signals. Based on the assumption that an aggregation of sentiment values on smaller time units is able to deliver more information as opposed to an aggregation on a longer time period, sentiment values were aggregated on a 10 minute time intervall. A smaller time intervall was also motivated by previous research as most studies were working with prices on higher aggregation levels, e.g., end-of-day prices in Crone and Koeppel (2014) or hourly price data in Papaioannou et al. (2013). In the following 13 sentiment-based indices were introduced to calculate binary trading signals (buy and sell). Based on every source three indices were generated, a positive index which only consisted of positive messages from a source, a negative index and a neutral index. In total 12 indices represented the number of positive, negative and neutral categories from different sources that were considered for data collection. The 13th index was created for control reasons and was a random index which was randomly comprised of values of the

other 12 indices. In order to trigger trading signals, sentiment values were generated based on moving medians for all 13 sentiment indices.

Analysis results showed that most indices exhibit a high degree of variability themselves and with respect to every other index. The order of integration of all sentiments was found to be 0 for all sentiment indices and 1 for the EUR/USD exchange rates, as evidenced by the augmented Dickey-Fuller (ADF) unit root test (e.g. Banerjee et al. 1993) for $p < 0.01$. These results indicated that sentiment indices are stationary while the EUR/USD exchange rate is not (Janetzko et al. 2017). Performance returns of most sentiment indices are characterized by high volatility and mean reversion which is a normal pattern of the random index. Especially, two indices, the positive and negative index of the source Stocktwits deviated from this pattern and showed the best performance compared to all other indices. The article further listed cumulative and cumulative abnormal returns for all sentiment indices.

While large differences were found for some sentiment indices further tests should reveal differential effects of buy and sell rules informed by sentiments. A closer look into separate buy and sell rules revealed that sell rules often perform better than buy rules. Calculation of skewness for all sentiment indices and for separate buy and sell rules also showed differences. If compared with random sell triggers the sell rules of the considered sentiment indices had a higher likelihood of triggering the sell signal in the right moment to be beneficial for the return of trading the EUR/USD exchange rate. Along with the added skewness values the result confirmed that the overall effect of the sell rule on returns was higher than that of the buy rule. This was echoed by higher skewness values of the returns gained with the sell rule when compared to the gains of the buy rule (Janetzko et al. 2017).

The main goal of Contribution V was to identify predictive value for different public online sources as previous research in this field lacked a comprehensive examination of this subject. Predictive value was measured by constructing a backtesting approach for trading the EUR/USD exchange rate. Comparing the results with a random index showed that the observed performance cannot be explained entirely by general market developments. The article therefore concluded that there is a strong indication for actual predictive value in the analyzed data. While some sources show large differences in performance it can be noted that sell signals seemed to have higher predictive value for the forex pair Euro/US Dollar. The best performing source was Stocktwits. Tweets from Stocktwits cleary showed higher predictive value in the analysis compared to messages from message boards or news websites. It might also be noteworthy that especially the positive and negative categories from Stocktwits showed very good performance results. Based on these findings it can be reasonably argued that neutral messages have lesser predictive value so this can be considered an expected result. The outperformance to tweets from Twitter which is very similar in nature to Stocktwits was explained with the assumption that on Twitter users have a lower degree of financial expertise compared to users of Stocktwits and therefore their contributions lead to a decreased usefulness for building financial forecasting models. Contribution V

can eventually provide evidence that methods of forecasting financial developments, which were successfully established for stock markets in previous research, can also be applied to currency trading.

7.8.2. Contribution and Position in Research Context

Contribution V addressed several limitations of Contribution IV as well as of other previous research from the same field. One major limitation of Contribution IV was the comparability of the predictive value of different data sources. The predictor was built on an aggregated sentiment score across all sources. Contribution V studied the predictive value of different types of sources of public online data using a unifying framework to achieve comparability of results. Furthermore, it used intraday data over a period of 3,5 years which addressed another limitation of Contribution IV that built the trading model on a rather short period of time (six months). According to figure 8 Contribution V is positioned in the area of sentiment analysis and predictive analytics.

As the majority of previous research was concerned with stocks and indices (e.g. Antweiler and Frank 2004, Bollen et al. 2011, Oh and Sheng 2011, Chen et al. 2014, Melville et al. 2009) Contribution V extended the research stream by studying sentiment based predictions for another kind of financial instrument: currencies. More precisely, data analysis in Contribution V was carried out with the goal of finding the most valuable type of source for predicting future exchange rates of EUR/USD. Studies by Papaioannou et al. (2013), Janetzko (2014), or Crone and Koeppel (2014) on predictive modeling of exchange rates provided evidence that sentiment indicators have predictive value for future exchange rate developments. In several regards, Contribution V went beyond these studies. Firstly, it extended the dataset studied from single to multi source (Papaioannou et al. 2013 and Janetzko 2014). Secondly, it set the focus on intraday data. Finally, Contribution V compared predictive values across different data sources (Papaioannou et al. 2013, Janetzko 2014, and Crone and Koeppel 2014) with an emphasis on robustness. Robustness of results was also a limitation of Contribution IV, mainly due to the short period of available historical data.

A more detailed view showed that the main deliverable of Contribution V, compared to previous research, was twofold: The setup of an analytical framework and the usage of a backtesting approach allowed to compare the effects of a range of different indices. It could be shown that some sentiment indices clearly outperform others. Above a pure effect analysis, Contribution V

(i) progressively narrowed down conditions that lead to increased performance which led to an uncovering of a dual effect of buy and sell trading rules and

(ii) rules out plausible alternative explanations (general market development)

Splitting up the overall effect of cumulative returns it became obvious that sell rules account for about two thirds while buy rules account for only one third of the overall effect. In contrast to other measures, e.g. skewness, returns were an extensive quantity that obeys the rule of additivity, and hence a splitting of the overall effects of cumulative returns was possible. Partial return effects were jointly generated by

a frequent application of a comparatively moderately effective rule (sell) and rare application of a relatively effective rule (buy).

As already the previous Contributions indicated for different domains, the results of Contribution V also suggested that data from publicly available online sources have the ability to predict future developments. In the case of Contribution V the findings showed that sentiment data from social media sources can improve trading models for the exchange rate EUR/USD. Further research will be necessary to find out whether these findings also apply to stocks and commodities trading and more importantly whether the found insights can eventually be generalized for financial markets.

8. Discussion and Conclusion

Signal or noise? This is the general question that has guided the work in all the contributions that comprise this dissertation. Do network structures harvested from social networking sites contain information that predicts the performance or success of groups or individuals? Insights on this question can be found in Contribution I (Krauss et al. 2008), Contribution II (Gloor et al. 2009), and Contribution III (Nann et al. 2009). The next section discusses these findings and the resulting contributions to the theory of (digital) social network analysis.

8.1. Contribution to Social Network Analysis

Contribution II (Gloor et al. 2009) offers a general model for measuring correlations between Web buzz and real-world effects. The main goal was to built a "Web Buzz Index" for different topics that enables prediction of the popularity of the topics in the future. The index uses centrality measures in combination with context-specific text mining approaches for publicly available data. It is a general derivative of the "Oscar Index" introduced in Contribution I (Krauss et al. 2008), which was constructed based on the network structures of users of a social movie forum and the communication content of text documents posted in that forum. The way in which the Web Buzz Index was built has not been widely adopted in prior research. In particular, the analysis of different public Web sources (websites, blogs, and forums) and the combination of network measures such as the concept of betweenness centrality and sentiment analysis approaches lead to a unique model that reveals relationships between Web buzz and real-world developments in different domains, such as stock markets. It has set the basis for further research in the field of predicting stock market developments (see Contributions IV, Nann et al. 2013, and V, Janetzko et al. 2017).

Contribution III (Nann et al. 2019) builds on the same methods to identify network structures and calculate centrality measures for a publicly available dataset collected from digital social networks. It shows that conclusions regarding the performance of groups or individuals in online social network can be reliably drawn by applying these methods. It further introduces a new concept for in-group networks, which are referred to as "tribal networks" or "tribes." The term "tribes" is defined as the ratio of the number of actors and edges within a given network to the number of actors and edges in an outside (external) network. It could be shown that high cohesiveness of tribes has a strong positive correlation with real-world effects, for example, on the economic impact of a university or on the impact of a company founder.

Contributions I (Krauss et al. 2008), II (Gloor et al. 2009), and III (Nann et al. 2009) use data from public domains. The newly developed methods are based on known centrality metrics from SNA and are combined with proprietary, context-specific sentiment analysis and text mining approaches. The findings of Contributions I, II, and III show that the application of these methods to different data from different public domains reveal predictive characteristics.

Do sentiment data extracted from UGC on digital social network sources have the ability to predict future financial market values? Contributions II (Gloor et al. 2009), IV (Nann et al. 2013), and V (Janetzko et al. 2017) provide answers to this question. Section 8.2 elaborates further on the individual contributions and also on the necessary relationship between Social Network Analysis, sentiment analysis, and predictive analytics.

8.2. Contribution to Predictive Analytics

The coexistence of SNA and sentiment analysis as described in the previous section can be adapted to this section as well. Sentiment analysis is a necessary antecedent for the application of models from predictive analytics. With sentiment analysis the tonality or emotions expressed in text documents can be extracted. The resulting emotional sentiment data was a refined resource which serves as input for models and methods applied in predictive analytics. Although, Antweiler and Frank (2004) were not the first ones who studied this area of research, they have provided a key article about the relationship of finance-related UGC and the development of stock prices. This study served as a basis for the construction of sentiment analysis approaches for all contributions of this dissertation except for Contribution III.

The approach of Antweiler and Frank (2004) has been extended in different ways. A key finding of Contribution IV (Nann et al. 2013) suggests that correct sentiment recognition depends on a context-specific application of the text mining methodologies built in a previous study by Krauss et al. (2012). A major contribution, therefore, is a specifically adapted sentiment analysis methodology to extract emotions from text documents with a focus on financial markets published in public social media sources. The consideration of different sources was another extension of previous research (Contribution V). In general, sentiment analysis and predictive analytics with a specific focus on financial markets have come a long way, with a strong increase in interest, especially in the first 10 years after the publication of Antweiler and Frank (2004) in the *Journal of Finance*. (See sections 4.2 and 6.1 for an overview of the development of articles in the field over time.)

The evaluation and aggregation stages through which obtained sentiment data pass have also been specifically developed and adapted for the contributions of this dissertation, and likewise extend methods used in previous research (section 6). Contribution IV (Nann et al. 2013) and Contribution V (Janetzko et al. 2017) make use of these methods in particular. The application of different concepts described in section 6 leads to an increased value and relevance of the sentiment data, as the results in Contributions IV and V show. For instance, proprietary spam detection and filtering methodologies enhance the quality of the data. Further, as various backtestings and applications in Contributions IV and V show, the inherent quality of the sentiment data is improved with consideration of a “half-life”-period of text documents and the application of an exponential smoothing factor to replicate real-world behavior as closely as possible.

Most prior research on predictive analytics is concerned with stocks and indices (e.g., Antweiler and Frank 2004, Bollen et al. 2011, Oh and Sheng 2011, Chen et al. 2014, Melville et al. 2009), whereas Contribution V (Janetzko et al. 2017) extends this stream by studying sentiment-based predictions for currencies. Another important contribution is the construction of a unifying backtesting framework to consider the predictive value of different types of public online sources for achieving comparability of results.

Contributions IV (Nann et al. 2013) and V (Janetzko et al. 2017), as well as parts of Contributon II (Gloor et al. 2009), provide strong indications that emotional sentiment data extracted from UGC and finance-related social media sources can predict future market values.

The methodologies constructed in this dissertation contribute significantly to research, especially to the theories of predictive analytics and social network analysis, and also add value in a practical environment. As described in section 5, a company (StockPulse) was founded based on major parts of and insights from the work presented herein. For several years, StockPulse has successfully applied sentiment analysis and predictive analytics methods to large-scale datasets mined from social media sources. Clients include global hedge funds, banks, stock exchanges, financial institutions, and financial publishers. From the author's perspective, the success of StockPulse and the practical relevance of the application of the methodologies partly developed within the contributions of this dissertation show a very high level of user validation.

8.3. Limitations

The datasets collected for the contributions, particularly those for Contributions IV and V that have been used for stock market analyses, do not follow a unifying scheme. For instance, the datasets differ in length (time period) and number of sources. Significant backtesting or trading models in financial markets usually require a length of at least 10 years, within which many different market phases will be present and it will be possible to observe correlations and patterns of the data during all phases. This enables comparability of different studies and eventually leads to robust models. However, as the field of research is comparatively new and the number of data sources has been growing exponentially over the last 10 years, it was not possible to create a common and consistent database across all contributions and calculations.

Findings based on data from online social networking sites should raise the question of robustness and meaning for the offline world. For instance, the dataset for Contribution III was collected for a restricted period of time and a specific subset of users of a special network site. To achieve generalizability, more data and different sources would be necessary. Further, qualitative analyses (e.g., offline questionnaires, surveys, etc.) should be executed in addition to the quantitative studies herein. However, as the number of people using online social networking sites or social media has been growing almost exponentially

over the last decade, it is only an assumption that the data inconsistency problem would be mitigated were the same analyses and methods to be applied to newly collected data.

8.4. Outlook

The use in the company StockPulse of the methodologies developed in and findings from the contributions of this dissertation show the potential for this field. The goal for the future, as new data sources are created constantly and the quantity of data continues to grow, is to develop general methods for using emotional sentiment data for prediction models in financal markets for all asset classes and any type of financial instrument. New approaches in artificial intelligence are currently developing that could take this potential even further, eventually supporting improved sentiment analysis and deliver better results in automatic text mining and emotion extraction.

Rapidly changing usage behavior of social media, especially among the younger generation, and an increased production of data in all areas of our lives will also lead to new challenges for research and the academic world. It will be necessary for researchers to understand the new ways of communication and the resulting influence these have on us all. Methods to analyze large-scale datasets in a quantitative way will become even more important tools for extracting insights. From the present author’s point of view, the methods and approaches developed in this dissertation contribute to achieving this goal.

References

General References

Ahuja, G. (2000). “Collaboration networks, structural holes, and innovation: A longitudinal study”. *Administrative Science Quarterly 45(3) 425–455*.

Banerjee, A.; Dolado, J.J.; Galbraith, J.W. and Hendry, D.F. (1993). “Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data”. *Oxford University Press, Oxford*.

Bavelas, A. (1950). Communication patterns in task oriented groups”. *Journal of the Acoustical Society of America 22:271-282.*

Bavelas, A. and Barrett D. (1951). “An experimental approach to organizational communication”. *Personnel 27:366-371*.

Blei, D.; Ng, A.Y. and Jordan, M.I. (2003). “Latent dirichlet allocation”. *Journal of machine Learning research 3(Jan):993–1022*

Brass, D.J. (1985). “Men’s and women’s networks: a study of interaction patterns and influence in an organization”. *Academy of Management Journal 28:327-43*.

Cummings, J. and Cross, R. (2003). “Structural properties of work groups and their consequences for performance”. *Social Networks, 25(3): 197-210*.

Daniel, K.; Hirshleifer, D. and Teoh, S.H. (2002). “Investor psychology in capital markets: evidence and policy implications”. *Journal of Monet Econ 49:139–209.*

Dellarocas, C.; Zhang, X. and Awad, N. F. (2007). “Exploring the value of online product reviews in forecasting sales: The case of motion pictures”. *Journal of Interactive Marketing, Volume 21, Issue 4, 2007, Pages 23-45.*

Dodge, M. and Kitchin, R. (2000). “Mapping Cyberspace”. *Routledge.*

Dodge, M. and Kitchin, R. (2002). “Atlas of Cyberspace”. *Pearson Education*.

Durkheim, E. (1960). “Sociology and Its Scientific Field”. *Pp. 354–75 in Emile Durkheim, 1858–1917: A Collection of Essays with Translations and a Bibliography, edited by K. H. Wolff. Columbus, OH: Ohio State University Press.*

Eliashberg, J. and Sawhney, M.S. (1996). “A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures”. *Marketing Science, 15 (2), 113-131*.

Freeman, L.C. (1979). “Centrality in Social Networks – Conceptual Clarification”. *In: Social Networks, Nr. 1, S. 215 – 239.*

Jensen, D. and Neville, J. (2002). “Data mining in social networks”. *Invited presentation to the National Academy of Sciences Workshop on Dynamic Social Network Modeling and Analysis, p. 7-9, Washington, DC.*

Malkiel, B.G. and Fama, E.F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work”. *The Journal of Finance 25(2):383-417, ISSN 1540-6261.*

Keynes, J.M. (1935). “The General theory of employment, interest and money”. *Macmillan Cambridge University Press, for Royal Economic Society.*

Nelson, R.A.; Donihue, M.R.; Waldman, D.M. and Wheaton, C. (2001). “What's an Oscar Worth?”. *Economic Inquiry, 39 (1).*

Podolny, J.M. and Baron, J.N. (1997). “Resources and relationships: social networks and mobility in the workplace”. *Am. Sociol. Rev. 62:673–93.*

Powell, W.W.; Koput, K.W. and Smith-Doerr, L. (1996). “Interorganizational collaboration and the locus of innovation”. *Administrative Science Quarterly 41(1): 116–145.*

Reagans, R. and Zuckerman, E.W. (2001). “Networks, diversity, and productivity: The social capital of corporate R&D teams”. *Organization Science, 12(4): 502–517.*

SalesChoice (2016). “Blog #2: A Closer Look at Artificial Intelligence”. *https://www.saleschoice.com/a-closer-look-at-artificial-intelligence/*

Simmel, G. (1950). “The Sociology of Georg Simmel, edited by K. H. Wolff”. *New York: The Free Press.*

Simonoff, J.S. and Sparrow, I.R. (2000). “Predicting movie grosses: Winners and losers, blockbusters and sleepers”. *Chance 13 (3), 15-24.*

Sparrowe, R.T.; Liden, R.C.; Wayne, S.J.; and Kraimer, M.L. (2001). “Social networks and the performance of individuals and groups”. *Academy of Management Journal, 44: 316 –325.*

Techopedia, *https://www.techopedia.com/definition/28799/data-intelligence, retrieved on June 10, 2019*

Tumarkin, R. and Whitelaw, R.F. (2001). “News or noise? internet postings and stock prices”. *Financial Analysts Journal, 57(3):41–51.*

Uzzi, B. (1996). “The sources and consequences of embeddedness for the economic performance of organizations: The network effect”. *American Sociological Review 61(4) 674–698.*

Wasserman, S. and Faust, K. (1994). “Social Network Analysis, Methods and Applications”. *Cambridge University Press.*

References Clustered By Years

2004

Adar, E.; Zhang, L.; Adamic, L. and Lukose, R. (2004). "Implicit Structure and the Dynamics of Blogspace: Workshop on the Weblogging Ecosystem". *13th International World Wide Web Conference.*

Antweiler, W. and Frank, M.Z. (2004). "Is all that talk just noise? The information content of internet stock message boards". *The Journal of Finance, 59 (3) : 1259–1294.*

Awad, N.F.; Dellarocas, C. and Zhang, X. (2004). "Is Online Word-of-mouth a Complement or Substitute to Traditional Means of Consumer Conversion". *Sixteenth Annual Workshop on Information Systems Economics (WISE), Washington, DC.*

Baker, M. and Stein, J. (2004). "Market liquidity as a sentiment indicator". *Journal of Financial Markets 7(3):271–299.*

Brown, G. and Cliff, M. (2004). "Investor sentiment and the near-term stock market". *Journal of Empirical Finance, May 13, 2002.*

Burt, R. (2004). "Structural Holes & Good Ideas". *American Journal of Sociology, (110): 349-99.*

Gloor, P. and Zhao, Y. (2004). "TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis, ACM CSCW Workshop on Social Networks". *ACM CSCW Conference, Chicago, 6. Nov. 2004.*

Lo, A.W. (2004). "The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective". *Journal of Portfolio Management: 30th Anniversary Issue, 15-29.*

Murray, F. (2004). "The role of academic inventors in entrepreneurial firms: Sharing the laboratory life". *Research Policy 33(4): 643–659.*

Qiu, L. and Welch, I. (2004). "Investor sentiment measures". *National Bureau of Economic Research.*

Surowiecki, J. (2004). "The wisdom of crowds". *Doubleday, New York.*

2005

Brown, G.W. and Cliff, M.T. (2005). "Investor sentiment and asset valuation". *The Journal of Business 78(2):405–440.*

Das, S.; Martinez-Jerez, A. and Tufano, P. (2005). "eInformation: A Clinical Study of Investor Discussion and Sentiment". *Financial Management 34(3):103-137.*

Hong, H.; Kubik J.D. and Stein, J.C. (2005). "Thy neighbor's portfolio: word-of-mouth effects in the holdings and trades of money managers". *Journal of Finance 60(6):2801–2824.*

Mayo, M. and Pastor, J.C. (2005). "Networks and effectiveness in work teams: the impact of diversity". *Instituto de Empresa Business School Working Paper No. WP05-10. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1016118. Accessed 2008-04-23*

Moran, P. (2005). "Structural vs. relational embeddedness: social capital and managerial performance". *Strategic Management Journal 26, 1129–1151.*

Nofsinger, J. (2005). "Social Mood and Financial Economics". *Journal of Behaviour Finance, S. 144-160.*

Pardoe, I. (2005). "Predicting Academy Award winners using discrete choice modeling". *In Proceedings of the 2005 Joint Statistical Meetings, Alexandria, VA. American Statistical Association.*

Whitelaw, C.; Garg, N. and Argamon, S. (2005). "Using Appraisal Groups for Sentiment analysis". *In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ACM, Bremen, Germany, pp. 625-631.*

2006

Al Hasan, M.; Chaoji, V.; Salem, S. and Mohammed, Z. (2006). "Link Prediction using Supervised Learning". *Proc 2006 Workshop on Link Analysis, Counterterrorism and Security.*

Baker, M. and Wurgler, J. (2006). "Investor sentiment and the cross-section of stock returns". *Journal of Finance, August, 2006.*

Balkundi, P. and D.A. Harrison. (2006). "Ties, Leaders, and Time in Teams: Strong Inference About Network Structure's Effects on Team Viability and Performance". *Academy of Management Journal 49(1): 49-68.*

Barber, B.M.; Lehavy, R.; McNichols, M. and Trueman, B. (2006). "Buys, holds, and sells: the distribution of investment banks' stock ratings and the implications for the profitability of analysts' recommendations". *J Acc Econ 41(1–2):87–117.*

Bono, J.E. and Ilies, R. (2006) "Charisma, positive emotions and mood contagion". *Leadersh Q 17:317–334.*

Bulkley, N. and Van Alstyne, M.W. (2006). "An Empirical Analysis of Strategies and Efficiencies in Social Networks". *MIT Sloan Research Paper No. 4682-08. Available at SSRN: http://ssrn.com/abstract=887406*

Chevalier, J.A. and Mayzlin, D. (2006). "The effect of word of mouth on sales: online book reviews". *J Mark Res 43(3):345–354.*

Kaplan, D. (2006). "And the Oscar Goes to… A Logistic Regression Model for Predicting Academy Award Results". *Journal of Applied Economics & Policy, 25 (1), 23-41.*

Lemmon, M. and Portniaguina, E. (2006). "Consumer confidence and asset prices: some empirical evidence". *Rev Financ Stud 19(4):1499–1529.*

Tseng, K.C. (2006). “Behavioral finance, bounded rationality, neurofinance, and traditional finance”. *Invest Manag Financ Innov 3(4):7–18.*

2007

Baker, M. and Wurgler, J. (2007). “Investor sentiment in the stock market”. *J Econ Perspect 21(2):129–151.*

Das, S. and Chen, M. (2007). “Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web”. *Management Science 53(9):1375-1388.*

Dellarocas, C.; Zhang, X. and Awad, N.F. (2007). „Exploring the value of online product reviews in forecasting sales: the case of motion pictures“. *Journal of Interact Mark 21(4):23–45.*

Edmans, A.; Garcia, D. and Norli, Ø. (2007). “Sports sentiment and stock returns”. *Journal of Finance 62(4):1967–1998.*

Java, A.; Song, X.; Finin, T. and Tseng, B. (2007). “Why We Twitter: Understanding Microblogging Usage and Communities”. *Joint 9th WEBKDD and 1st SNA-KDD Workshop. San Jose, CA.*

Ku, L.W. and Chen, H.H. (2007). “Mining Opinions from the Web: Beyond Relevance Retrieval”. *J. American Society for Information Science and Technology, vol. 58, no.12, 2007, pp. 1838–1850.*

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N. and Hurst, M. (2007). “Cascading behavior in large blog graphs”. *Proceedings of the 7th SIAM international conference data mining (SDM).*

Malmendier, U. and Shanthikumar, D. (2007). “Are small investors naive about incentives?”. *Journal of Finance Econ 85(2):457–489.*

Mitchell, M.L. and Mulherin, J.H. (2007). “The impact of public information on the stock market”. *Journal of Finance 49(3):923–950.*

Raz, O. and Gloor, P. (2007). “Size Really Matters - New Insights for Startup's Survival”. *Management Science*

Tetlock, P. (2007). “Giving content to investor sentiment: The role of media in the stock market”. *Journal of Finance 62:1139-1168.*

Vickery, G. and Wunsch-Vincent, S. (2007). “Participative Web and User-Created Content: Web 2.0”. *Wikis and Social Networking. http://browse.oecdbookshop.org/oecd/pdfs/free/9307031e.pdf, retrieved 12/31/2013.*

2008

Abbasi, A.; Chen, H. and Salem, A. (2008). “Sentiment analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums”. *ACM Trans. Inf. Syst. (26:3) 1-34.*

Barber, B.M. and Odean, T. (2008). "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". *Review of Financial Studies 21(2): 785–818.*

Chang, S.C.; Chen, S.S.; Chou, R.K. and Lin, Y.H. (2008). "Weather and intraday patterns in stock returns and trading activity". *J Bank Financ 32:1754–1756.*

Cohen, L. and Frazzini, A. (2008). "Economic Links and Predictable Returns". *Journal of Finance 63:1977-2011.*

De Choudhury, M.; Sundaram, H.; John, A. and Seligmann, D.D. (2008). "Can blog communication dynamics be correlated with stock market activity?". *In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, pages 55–60. ACM.*

Fowler, J.H. and Christakis, N.A. (2008). "Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years". *Framingham Heart Study. BMJ 337:a2338*

Hancock, J.T.; Gee, K.; Ciaccio, K. and Lin, J.M.H. (2008). "I'm sad you're sad: emotional contagion in CMC". *Proceedings of the 2008 ACM conference computing supported cooperative work.*

Hinz, O. and Spann, M. (2008). "The impact of information diffusion on bidding behavior in secret reserve price auctions". *Information System Research 19(3):351–368.*

Howe, J. (2008). "Crowdsourcing: why the power of the crowd is driving the future of business". *Crown Business, New York.*

Kittur, A. and Kraut, R.E. (2008). "Harnessing the wisdom of crowds in Wikipedia: quality through coordination". *In: Proceedings of the 2008 ACM conference on Computer supported cooperative work, pp 37–46.*

Kleinberg, J. (2008). "The Convergence of Social and Technological Networks". *Communications of the ACM, Vol. 51 No. 11 November 2008, pp. 66-72.*

Krauss, J.; Nann, S.; Simon, D.; Fischbach, K. and Gloor, P.A. (2008). "Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis". *16th European Conference on Information Systems (ECIS).*

Mayer, A. and Puller, A. (2008). "The old boy (and girl) network: Social network formation on university campuses". *Journal of Public Economics, Volume 92, Issues 1-2, February 2008, Pages 329-347.*

Pang, B. and Lee, L. (2008). "Opinion mining and sentiment analysis". *Foundations and Trends in Information Retrieval 2 (2008) 1–135.*

Sabherwal, S.; Sarkar, S. and Zhang, Y. (2008). "Online talk: does it matter?". *Managerial Finance 34(6):423-436.*

Tetlock, P.C.; Saar-Tsechansky, M. and Macskassy, S. (2008). "More than words: Quantifying language to measure firms' fundamentals". *The Journal of Finance 63(3): 1437–1467.*

2009

Albuquerque, R. and Vega, C. (2009). "Economic news and international stock market co-movement". *Rev. Finance, vol. 13, no. 3, pp. 401–465.*

Boiy, E. and Moens, M. (2009). "A Machine Learning Approach to Sentiment analysis". *In Multilingual Web Texts, Information Retrieval (12:5) 526.*

Choi, Y.; Kim, Y. and Myaeng, S.H. (2009). "Domain-Specific Sentiment analysis Using Contextual Feature Generation". *In Proceeding of the 1st international CIKM Workshop on Topic-sentiment analysis for Mass Opinion, ACM, Hong Kong, China, pp. 37-44.*

Dhar, V. and Chang, E. (2009). "Does chatter matter? The impact of user-generated content on music sales". *J Interact Mark 23(4):300–307.*

Doshi, L.; Krauss, J.; Nann, S. and Gloor, P.A. (2009). "Predicting Movie Prices Through Dynamic Social Network Analysis". *Proceedings of Collaborative Innovations Networks Conference COINs*

Fang, L. and Peress, J. (2009). "Media coverage and the cross-section of stock returns". *Journal of Finance, vol. 64, no. 5, pp. 2023–2052.*

Gloor, P.A.; Krauss, J.; Nann, S.; Fischbach, K. and Schoder, D. (2009). "Web science 2.0: Identifying trends through semantic social network analysis". *Computational Science and Engineering, 2009. CSE'09. International Conference on. Vol. 4. IEEE, 2009.*

Hey, T.; Tansley, S. and Tolle, K. (2009). "The Fourth Paradigm – Data-Intensive Scientific Discovery". *Microsoft Research, Redmond, Washington.*

Lin, C. and He, Y. (2009). "Joint Sentiment/Topic Model for Sentiment analysis". *In Proceeding of the 18th ACM conference on Information and Knowledge Management, ACM, Hong Kong, China, pp. 375-384.*

Melville, P.; Gryc, W. and Lawrence, R.D. (2009). "Sentiment analysis of blogs by combining lexical knowledge with text classification". *In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1275-1284.*

Mizrach, B. and Weerts, S. (2009). "Experts online: An analysis of trading activity in a public Internet chat room". *Journal of Economic Behavior and Organization 70(1): 266–281.*

Nann, S.; Krauss, J.; Schober, M.; Gloor, P.A.; Fischbach, K. and Führes, H. (2009). "Comparing the structure of virtual entrepreneur networks with business effectiveness". *Proceedings of Collaborative Innovations Networks Conference (COINs), Savannah, GA.*

Narayanan, R.; Liu, B. and Choudhary, A. (2009). "Sentiment analysis of conditional sentences". *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, Association for Computational Linguistics, Singapore, pp. 180-189.*

Schumaker, R. and Chen, H. (2009). "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System". *ACM Transactions on Information Systems 27(2).*

Schumaker, R.P. and Chen, H. (2009). "A quantitative stock prediction system based on financial news". *Inf. Process. Manag., vol. 45, no. 5, pp. 571–583.*

Worthington, A. (2009). "An empirical note on weather effects in the Australian stock market". *Econ Pap J Appl Econ Policy 28(2):148–154.*

Xu, S. and Zhang, X. (2009). "How do social media shape the information environment in the financial market?". *International Conference on Information Systems (ICIS) 2009.*

Zeledon, M. (2009). "StockTwits may change how you trade". *New York: Bloomberg Businessweek. http://www.businessweek.com/technology/content/feb2009/tc20090210_875439.htm*

2010

Asur, S. and Huberman, B.A. (2010). "Predicting the Future with Social Media". *WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, Pages 492-499.*

Bifet, A. and Frank, E. (2010). "Sentiment knowledge discovery in Twitter streaming data". *Discov Sci 6332:1–15.*

Bollen, J.; Pepe, A. and Mao, H. (2010). "Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena". *Proceedings from 19th International World Wide Web Conference Raleigh, North Carolina.*

Boyd, D.; Golder, S. and Lotan, G. (2010). "Tweet, tweet, retweet: conversational aspects of retweeting on twitter". *Proceedings of the 43rd Hawaii international conference on social systems (HICSS).*

Cha, M.; Haddadi, H.; Benevenuto, F. and Gummadi, K.P. (2010). "Measuring user influence in Twitter: the million follower fallacy". *Proceedings of the 4th international aaai conference on weblogs and social media.*

Chintagunta, P.K.; Gopinath, S. and Venkataraman, S. (2010). "The effects of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets". *Marketing Science 29(5):944–957.*

Gilbert, E., Karahalios, K. (2010). "Widespread worry and the stock market". *Proceedings of the 4th international AAAI conference weblogs social media.*

Gloor, P.; Oster, D.; Raz, O.; Pentland, A. and Schoder, D. (2010). "The Virtual Mirror - Reflecting on Your Social and Psychological Self to Increase Organizational Creativity". *Journal of International Studies of Management & Organization Volume 40, 2010 - Issue 2, 74-94.*

Kwak, H.; Lee, C.; Park, H. and Moon, S. (2010). "What is Twitter, a social network or a news media?". *Proceedings of the 19th international conference world wide web.*

Schumaker, R.P. and Chen, H. (2010) "A discrete stock price prediction engine based on financial news". *Computer, vol. 43, no. 1, pp. 51–56.*

Sprenger, T.O. and Welpe, I.M. (2010). "Tweets and Trades: The Information Content of Stock Microblogs". *Working Paper, http://ssrn.com/abstract=1702854, retrieved 04/22/2012.*

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H. and Demirbas, M. (2010). "Short text classification in twitter to improve information filtering". *In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval (SIGIR 2010), Geneva, Switzerland, July 19–23 (pp. 841–842). ACM.*

Tetlock, P.C. (2010). "Does public financial news resolve asymmetric information?". *Rev. Financ. Stud., vol. 23, no. 9, pp. 3520–3557.*

Yang, J. and Counts, S. (2010). "Predicting the speed, scale, and range of information diffusion in Twitter". *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (vol. 10, pp. 355–358). Washington, DC: Association for the Advancement of Artificial Intelligence.*

Ye, S. and Wu, S.F. (2010). "Measuring message propagation and social influence on twitter.com". *Soc Inform 6430:216–231.*

Yerva, S.R.; Miklós, Z. and Aberer, K. (2010). "It was easy, when apples and blackberries were only fruits". *EPFL working paper, http://infoscience.epfl.ch/record/151616/files/LSIR_WePS3_Paper.pdf, retrieved 03/30/2013.*

Zhang, W. and Skiena, S. (2010). "Trading strategies to exploit blog and news sentiment". *Proceedings of 4th Int. AAAI Conf. Weblogs and Social Media, 2010, pp. 375– 378.*

Zhang, X.; Fuehres, H. and Gloor, P. (2010). "Predicting stock market indicators through Twitter – ''I hope it is not as bad as I fear''". *Proceedings of Collaborative Innovations Networks Conference (COINs), Savannah, GA.*

Zhang, Y. and Swanson, P.E. (2010). "Are day traders bias free? Evidence from Internet stock message boards". *Journal of Economics and Finance 34(1): 96–112.*

Zhao, J.M.; He, X. and Wu, F.Y. (2010). "Study on Chinese stock market rumors and the impact on stock prices". *Manag. World, vol. 11, pp. 38–51.*

Zhu, F. and Zhang, X. (2010). "Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics". *Journal of Marketing 74(2):133–148.*

2011

Bakshy, E.; Hofman, J.M.; Mason, W.A. and Watts, D.J. (2011). "Everyone's an influencer: quantifying influence on Twitter". *Proceedings 4th ACM international conference web search data, pp 65–74.*

Birz, G. and Lott, J.R. (2011). "The effect of macroeconomic news on stock returns: New evidence from newspaper coverage". *Journal of Bank. Finance, vol. 35, no. 11, pp. 2791–2800.*

Bollen, J. and Mao, H. (2011). "Twitter Mood As a Stock Market Predictor". *Computer 44(10):91-94, ISSN 0018-9162.*

Bollen, J.; Mao, H. and Zeng, X. (2011). "Twitter mood predicts the stock market" *Journal of Computational Science 2(1): 1–8.*

Carretta, A.; Farina, V.; Martelli, D.; Fiordelisi, F. and Schwizer, P. (2011). "The impact of corporate governance press news on stock market returns". *Eur. Financ. Manag., vol. 17, no. 1, pp. 100–119.*

Chan, S.W. and Franklin, J. (2011). "A text-based decision support system for financial sequence prediction". *Decision Support Systems 52 (2011) 189–198.*

Chen, H.; Huang, E.C.N.; Lu, H.M. and Li, S.H. (2011). "AZ SmartStock: Stock prediction with targeted sentiment and life support". *IEEE Intell. Syst, vol. 26, no. 6, pp. 84–88, Nov.-Dec. 2011.*

Dion, M.; Shaikh, V.; Pessaris, A.; Malin, S.; Smith, R.; Lakos-Bujas, D.; Okamoto, S. and Hlavaty, B. (2011). "News Analytics - Can They Add Value to Your Quant Process? – Using a Language Recognition Algorithm to Analyse News Flow". *J.P. Morgan Europe Equity Research, June 3rd, 2011.*

Engelberg, J.E. and Parsons, C.A. (2011). "The causal impact of media in financial markets". *Journal of Finance, vol. 66, no. 1, pp. 67–97.*

Ghosh, R., and Lerman, K. (2011). "A framework for quantitative analysis of cascades on networks". *Proceedings of the 4th ACM international conference web search data mining.*

Griffin, J.M.; Hirschey, N.H. and Kelly, P.J. (2011). "How important is the financial media in global markets?". *Rev. Financ. Stud., vol. 24, pp. 3941–3992.*

Groß-Klußmann, A. and Hautsch, N. (2011). "When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions". *Journal of Empirical Finance, vol. 18, no. 2, pp. 321–340.*

Groth, S.S. and Muntermann, J. (2011). "An intraday market risk management approach based on textual analysis". *Decision Support Systems 50 (2011) 680–691.*

Guillory, J.; Spiegel, J.; Drislane, M.; Weiss, B.; Donner, W. and Hancock, J. (2011). "Upset now?: emotion contagion in distributed groups". *Proceedings of the SIGCHI conference human factors computing system.*

Hill, S. and Ready-Campbell, N. (2011). “Expert stock picker: the wisdom of (experts in) crowds”. *Int Journal of Electron Commer 15(3):73–102.*

Hinz, O.; Skiera, B.; Barrot, C. and Becker, J. (2011). “Seeding strategies for viral marketing: an empirical comparison”. *Journal of Marketing 75(6):55–71.*

Kara, Y.; Boyacioglu, M.A. and Baykan, Ö.K. (2011). “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange”. *Expert Syst. Appl., vol. 38, no. 5, pp. 5311–5319.*

Karabulut, Y. (2011). “Can Facebook predict stock market activity?”. *Working Paper, University of Frankfurt, Germany.*

Levy, T., Yagil, J. (2011). “Air pollution and stock returns in the US”. *Journal of Econ Psychol 32(3):374–383.*

Mao, H.; Counts, S. and Bollen, J. (2011). “Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data“. *arXiv:1112.1051 URL http://arxiv.org/abs/1112.1051*

Oh, C., and Sheng, O. (2011). “Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement”. *Thirty Second International Conference on Information Systems (ICIS), Shanghai, pp 1-19.*

Shmueli, G. and Koppius O.R. (2011). “Predictive analytics”. *Information Systems Research MIS*

Tetlock, P.C. (2011). “All the news that’s fit to reprint: Do investors react to stale information?”. *Review of Financial Studies, 24(5): 1481–1512.*

Wang, B.; Huang, H. and Wang, X. (2011). “A novel text mining approach to financial time series forecasting”. *Neurocomputing, vol. 83, pp. 136–145.*

2012

Brown, E.D. (2012). “Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market”. *SAIS 2012 Proceedings. Paper 7.*

Chang, S.C.; Chen, S.S.; Chou, R.K. and Lin, Y.H. (2012). “Local sports sentiment and returns of locally headquartered stocks: a firmlevel analysis”. *Journal of Empirical Finance 19(3):309–318.*

Fang, F.; Datta, A. and Dutta, K. (2012). “A Hybrid Method for Cross-domain Sentiment Classification Using Multiple Sources”. *ICIS 2012, Orlando, Florida, USA.*

Giannini, R.; Irvine, P. and Shu, T. (2012). “The Convergence and Divergence of Investors’ Opinions around Earnings News: Evidence from a Social Network”. *Asian Finance Association (AsFA) 2013 Conference, March 8, 2012.*

Heimbach, I. and Hinz, O. (2012). "How smartphone apps can help predicting music sales". *Proceedings of the 20th European conference information system (ECIS), Barcelona, Spain.*

Hertwig, R. (2012). "Tapping into the wisdom of the crowd—with confidence". *Science 336:303–304.*

Kramer, A.D. (2012). "The spread of emotion via facebook". *Proceedings of the SIGCHI conference human factors computer system, pp 767–770.*

Krauss, J.; Nann, S. and Schoder, D. (2012). "Towards Universal Sentiment analysis through Web Mining". *Poster Session, European Conference on Information Systems (ECIS), Barcelona, Spain.*

Lerman, K.; Ghosh, R. and Surachawala, T. (2012). "Social contagion: an empirical study of information spread on Digg and Twitter follower graphs". *arXiv:1202.3162.*

Lu, R.; Xu, Z., Zhang, Y. and Yang, Q. (2012). "Life activity modeling of news event on twitter using energy function". *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II (PAKDD'12).*

Mizumoto, K.; Yanagimoto, H. and Yoshioka, M. (2012). "Sentiment analysis of Stock Market News with Semi-supervised Learning". *Proceedings of the 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, 325-328.*

Nizer, P. and Nievola, J.C. (2012). "Predicting published news effect in the Brazilian stock market". *Expert Syst. Appl., vol. 39, no. 12, pp. 10 674– 10 680.*

Rao, T. and Srivastava, S. (2012). "Using Twitter sentiments and search volumes index to predict oil, gold, forex and markets indices". *https://repository.iiitd.edu.in/jspui/handle/123456789/31*

Ruiz, E.J.; Hristidis, V.; Castillo, C.; Gionis, A. and Jaimes, A. (2012). "Correlating financial time series with micro-blogging activity". *WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining, Pages 513-522*

Schumaker, R.; Zhang, Y.; Huang, C.N. and Chen, H. (2012). "Evaluating sentiment in financial news articles". *Decision Support Systems, Volume 53, Issue 3, June 2012, Pages 458–464.*

Tirunillai, S. and Tellis, G.J. (2012). "Does chatter really matter? Dynamics of user-generated content and stock performance". *Marketing Science 31(2):198–215.*

Valerius, P. (2012). "Opinion Mining - Nutzung von Twitter als Meinungsquelle zur Vorhersage von Börsenkursen". *Universität Koblenz.*

Vu, T.T.; Chang, S.; Ha, Q.T. and Collier, N. (2012). "An experiment in integrating sentiment features for tech stock prediction in twitter". *Proceedings. Workshop Inform. Extraction Entity Analytics Social Media Data, 2012, pp. 23–38.*

Wang, S.; Luo, Y.; Cahan, R.; Alvarez, M.; Jussa, J. and Chen, Z. (2012). "Signal Processing: The Rise of the Machines". *Deutsche Bank Quantitative Strategy, 5 June 2012.*

Xu, W.; Li, T.; Jiang, B. and Cheng, C. (2012). "Web Mining For Financial Market Prediction Based On Online Sentiments". *PACIS 2012 Proceedings. Paper 43.*

2013

Alanyali, M.; Moat, H.S. and Preis, T. (2013). "Quantifying the relationship between financial news and the stock market". *Sci. Rep., vol. 3, pp. 1–6.*

D'Onfro, J. (2013). "Twitter admits 5 of its 'users' are fake". *http://www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10.*

Egger, M. and Lang, A. (2013). "A Brief Tutorial on How to Extract Information from User-Generated Content (UGC)". *KI - Künstliche Intelligenz, Volume 27, Issue 1, pp 53-60.*

Erikson, E. (2013). "Formalist and relationalist theory in social network analysis". *Sociological Theory, 31: 219-242.*

Gagnon, S. (2013). "Rules-based integration of news-trading algorithm". *Journal of Trading 8 (2013) 15–27.*

Hagenau, M.; Hauser, M.; Liebmann, M. and Neuman,n D. (2013). "Reading all the news at the same time: Predicting mid-term stock price developments based on news momentum". *In: 46th Hawaii International Conference on System Sciences (HICSS), 2013, pp. 1279–1288.*

Hagenau, M.; Liebmann, M. and Neumann, D. (2013). "Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decis. Support Syst., vol. 55, no. 3, pp. 685– 697.*

Lee, E. (2013). "AP Twitter account hacked in market-moving attack". *Bloomberg. https://www.bloomberg.com/news/articles/2013-04-23/dow-jones-drops-recovers-after-false-report-on-ap-twitter-page. Accessed 24 April 2013.*

Li, Y.M. and Li, T.Y. (2013). "Deriving market intelligence from microblogs". *Decision Support Systems 55 (2013) 206–217.*

Loughlin, C. and Harnisch, E. (2013). "The viability of StockTwits and Google Trends to predict the stock market". *At http://stocktwits.com/research/Viability-of-StockTwits-and-Google-Trends-Loughlin_Harnisch.pdf, accessed 12 May 2014.*

Luo, X., Zhang, J. and Duan, W. (2013). "Social media and firm equity value". *Information Systems Research 24(1): 146–163.*

Luo, Y.; Wang, S.; Cahan, R.; Jussa, J:; Chen, Z. and Alvarez, M. (2013). "DB Handbook of Portfolio Construction, Part I". *Deutsche Bank Quantitative Strategy, May 30, 2013.*

Makrehchi, M.; Shah, S. and Liao, W. (2013). "Stock prediction using event-based sentiment analysis". *Proceedings of IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., 2013, pp. 337 –342.*

Moat, H.S.; Curme, C.; Avakian, A.; Kenett, D.Y.; Stanley, H.E. and Preis, T. (2013). "Quantifying Wikipedia usage patterns before stock market moves". *Scientific Reports volume 3, Article number: 1801, pp. 1 –5.*

Nann, S.; Krauss, J. and Schoder, D. (2013). "Predictive Power on Public Data – the Case of Stock Markets". *Proceedings of 21st European Conference on Information Systems (ECIS).*

Oliveira, N.; Cortez, P. and Areal, N. (2013). "On the predictability of stock market behavior using stocktwits sentiment and posting volume". *In Correia L., Reis L.P., Cascalho J. (eds) Progress in Artificial Intelligence. EPIA 2013. Lecture Notes in Computer Science, vol 8154. Springer, Berlin, Heidelberg, pp. 355-365.*

Papaioannou, P.; Russo, L.; Papaioannou,G. and Siettos, C.I. (2013). "Can social microblogging be used to forecast intraday exchange rates". *Netnomics: Economic Research And Electronic Networking, (14:1-2), pp. 47–68.*

Preis, T.; Moat, H.S. and Stanley, H.E. (2013). "Quantifying trading behavior in financial markets using Google trends". *Science Reports, vol. 3, pp. 1–6.*

Siering, M. (2013). "Investigating the impact of media sentiment and investor attention on financial markets". *In: Lecture Notes in Business Information Processing,volume 135, Springer, Berlin, Heidelberg, 2013, pp. 3–19.*

US Securities and Exchange Commission (2013). "SEC Says Social Media OK for Company Announcements if Investors Are Alerted". *Online available at http://www.sec.gov/News/PressRelease/Detail/PressRelease/1365171513574.*

Xu, S.X. and Zhang, X.M. (2013). "Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction". *MIS Quarterly 37(4): 1043–1068.*

Yu, Y.; Duan, W. and Cao, Q. (2013). "The impact of social and conventional media on firm equity value: A sentiment analysis approach". *Decision Support Systems 55(4): 919–926.*

2014

Al Nasseri, A.; Tucker, A. and de Cesare, S. (2014). "Big Data Analysis of StockTwits to Predict Sentiments in the Stock Market". *Džeroski, S.; Panov, P.; Kocev, D.; Todorovski, L. (eds) Discovery Science. DS 2014. Lecture Notes in Computer Science, vol 8777. Springer, Cham.*

Ammann, M.; Frey, R. and Verhofen, M. (2014). "Do newspaper articles predict aggregate stock returns?". *J. Behav. Finance, vol. 15, no. 3, pp. 195–213.*

Chen, H.; De, P.; Hu, Y. and Hwang, B.H. (2014). "Wisdom of crowds: The value of stock opinions transmitted through social media". *The Review of Financial Studies, 27(5), pp. 1367-1403.*

Christakis, N.A. and Fowler, J.H. (2014). “Detecting emotional contagion in massive social networks”. *PLoS One 9(3):e90315*

Coviello, L.; Sohn, Y.; Kramer, A.D.; Marlow, C.; Franceschetti, M.; Christakis, N.A. and Fowler, J.H. (2014). “Detecting emotional contagion in massive social networks”. *PLoS One 9(3):e90315.*

Crone, S. F. and Koeppel, C. (2014). “Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons”. *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2104 IEEE Conference on, IEEE, London, UK, pp. 114-121.*

Curme, C.; Preis, T.; Stanley, H.E. and Moat, H.S. (2014). “Quantifying the semantics of search behavior before stock market moves”. *Proceedings of Natl. Acad. Sci., vol. 111, no. 32, pp. 11 600–11 605.*

de Fortuny, E.J.; de Smedt, T.; Martens, D. and Daelemans, W. (2014). “Evaluating and understanding text-based stock price prediction models”. *Information Processing & Management 50 (2014) 426–441.*

Ding, X.; Zhang, Y.; Liu, T. and Duan, J. (2014). “Using structured events to predict stock price movement: An empirical investigation”. *Proceedings of Conference of Empirical Methods Natural Language Process., 2014, pp. 1415–1425.*

Geva, T. and Zahavi, J. (2014). “Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news”. *Decision Support Systems 57 (2014) 212–223.*

Gottschlich, J. and Hinz, O. (2014). “A decision support system for stock investment recommendations using collective wisdom”. *Decision Support Systems Volume 59, March 2014, Pages 52-62.*

Groth, S.S.; Siering, M. and Gomber, P. (2014). “How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data”. *Decision Support Systems 62 (2014) 32–42.*

Hentschel, M. and Alonso, O. (2014). “Follow the money: A study of cashtags on Twitter”. *First Monday, [S.l.], aug. 2014. ISSN 13960466. Available at: <https://journals.uic.edu/ojs/index.php/fm/article/view/5385>. Date accessed: 22 june 2019.*

IHS Markit Research Signals (2014). “Alpha: Extracting Market Sentiment From 140 Characters”. March 3, 2014.

Janetzko, D. (2014). “Predictive modeling in turbulent times–What Twitter reveals about the EUR/USD exchange rate”. *Netnomics: Economic Research and Electronic Networking (15:2), pp. 69-106.*

Jiang, C.; Liang, K.; Chen, H. and Ding, Y. (2014). "Analyzing market performance via social media: A case study of a banking industry crisis". *Sci. China Inf. Sci., vol. 57, no. 5, pp. 1–18.*

Kearney, C. and Liu, S. (2014). "Textual sentiment in finance: A survey of methods and models". *International Review of Financial Analysis 33 (2014) 171–185.*

Kim, S.H. and Kim, D. (2014). "Investor sentiment from internet message postings and the predictability of stock returns". *Journal of Econ. Behav. Organ., vol. 107, pp. 708 –729.*

Kramer, A.D.; Guillory, J.E. and Hancock, J.T. (2014). "Experimental evidence of massive-scale emotional contagion through social networks". *Proceedings of the National Academy of Science.*

Li, Q.; Wang, T.; Gong, Q.; Chen, Y.; Lin, Z. and Song, S. (2014). "Media-aware quantitative trading based on public web information". *Decision Support Systems Volume 61, May 2014, Pages 93-105*

Li, Q.; Wang, T.; Li, P.; Liu, L.; Gong, Q. and Chen, Y. (2014). "The effect of news and public mood on stock movements": *Inf. Sci., vol. 278, pp. 826–840.*

Li, X.; Huang, X.; Deng, X. and Zhu , S. (2014). "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information". *Neurocomputing, vol. 142, pp. 228–238.*

Li, X.; Xie, H.; Chen, L.; Wang, J. and Deng, X. (2014). "News impact on stock price return via sentiment analysis". *Knowl.-Based Syst., vol. 69, pp. 14–23.*

Li, Z.; Xu, W.; Zhang, L. and Lau, R.Y. (2014). "An ontology-based web mining method for unemployment rate prediction". *Decision Support Systems 66 (2014) 114–122.*

Luo, Y.; Wang, S.; Alvarez, M.; Jussa, J.; Wang, A. and Rohal, G. (2014). "Signal Processing: Macro Uncertainty, Investor Sentiment, and Asset Returns". *Deutsche Bank Quantitative Strategy, September 15, 2014.*

Marshall, B.R.; Visaltanachoti, N. and Cooper, G. (2014). "Sell the rumour, buy the fact?". *Account. Finance, vol. 54 , no. 1, pp. 237–249.*

Medhat, W.; Hassan, A. and Korashy, H. (2014). "Sentiment analysis algorithms and applications: survey". *Ain Shams Engineering Journal, 5(4), pp.1093-1113.*

Nassirtouss, A.K.; Aghabozorgi, S.; Wah, T.Y. and Ngo, D.C.L. (2014). "Text mining for market prediction: A systematic review". *Expert Syst. Appl., vol. 41, no. 16, pp. 7653–7670.*

Nofer, M. and Hinz, O. (2014). "Are Crowds on the Internet Wiser than Experts? - The Case of a Stock Prediction Community". *Journal of Business Economics, 84 (3), 303-338.*

Peress, J. (2014). "The media and the diffusion of information in financial markets: Evidence from newspaper strikes". *Journal of Finance, vol. 69, no. 5, pp. 2007–2043.*

Rao, T. and Srivastava, S. (2014). “Twitter sentiment analysis: How to hedge your bets in the stock markets”. *In Fazli Can, Tansel Özyer & Faruk Polat (Eds), State of the Art Applications of Social Network Analysis (227-247). Springer International Publishing.*

Si, J.; Mukherjee, A.; Liu, B.; Pan, J.S.; Li, Q. and Li, H. (2014). “Exploiting social relations and sentiment for stock prediction”. *Proceedings of Conference of Empirical Methods Natural Language Process., 2014, pp. 1139–1145.*

Sprenger, T.O.; Tumasjan, A.; Sandner, P.G. and Welpe, I.M. (2014) “Tweets and Trades: the Information Content of Stock Microblogs”. *European Financial Management 20(5):926-957, ISSN 1468-036X.*

Yang, X. and Luo, Y. (2014). “Rumor clarification and stock returns: Do bull markets behave differently from bear markets?”. *Emerg. Markets Finance Trade, vol. 50, no. 1, pp. 197–209.*

Zheludev, I.; Smith, R. and Aste, T. (2014). “When can social media lead financial markets?” *Science Reports, vol. 4, pp. 1–12.*

2015

Ahern, K.R. and Sosyura, D. (2015). “Rumor has it: Sensationalism in financial media”. *Rev. Financ. Stud., vol. 28, pp. 2050–2093.*

Blaschak, J.; Blinov, A.; Gits, J.A.; Harfoush, F. and Myers, K. (2015). “Systems and Methods of Detecting, Measuring, and Extracting Signatures of Signals Embedded in Social Media Data Streams”. *U.S. Patent No 20180373703, 27.12.2018.*

Crosby, P. (2015). “Crowd-Sourced Stock Sentiment Using StockTwits.”, *Online Available at https://www.quantopian.com/posts/crowd-sourced-stock-sentiment-using-stocktwits*.

Ding, X.; Zhang, Y.; Liu, T. and Duan, J. (2015). “Deep learning for event-driven stock prediction”. *Proceedings of 24th Int. Joint Conf. Artificial Intell., pp. 2327–2333.*

Lillo, F.; Miccichè, S.; Tumminello, M.; Piilo, J. and Mantegna, R.N. (2015). “How news affects the trading behaviour of different categories of investors in a financial market”. *Quant. Finance, vol. 15, no. 2, pp. 213–229.*

Liu, L.; Wu, J.; Li, P. and Li, Q. (2015). “A social-media-based approach to predicting stock comovement”. *Expert Syst. Appl., vol. 42, no. 8, pp. 3893–3901.*

Luss, R. and d’Aspremont, A. (2015), “Predicting abnormal returns from news using text classification”. *Quant. Finance, vol. 15, no. 6, pp. 999–1012.*

Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y. and Ngo, D.C.L. (2015). “Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment”. *Expert Syst. Appl., vol. 42, no. 1, pp. 306 –324.*

Nguyen, T.H.; Shirai, K. and Velcin, J. (2015). “Sentiment analysis on social media for stock movement prediction”. *Expert Syst. Appl., vol. 42, no. 24, pp. 9603–9611.*

Nofer, M. and Hinz, O. (2015). “Using Twitter to Predict the Stock Market Where is the Mood Effect?”. *Bus Inf Syst Eng 57(4):229–242*

Ranco, G.; Aleksovski, D.; Caldarelli, G.; Grčar, M. and Mozetič, I. (2015). “The effects of Twitter sentiment on stock price returns”. *PLoS ONE 10(9): e0138441.*

Ravi, K. and Ravi, V. (2015). “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications”. *Knowledge-Based Systems 89 (2015) 14–46.*

Serrano-Guerrero, J.; Olivas, J.A.; Romero, F.P. and Herrera-Viedma, E. (2015). “Sentiment analysis: A review and comparative analysis of web services”. *Information Sciences Volume 311, 1 August 2015, Pages 18-38*

Thakkar, H. and Patel, D. (2015). “Approaches for Sentiment analysis on Twitter: A State-of-Art study”. *CoRR.*

Yang, S.Y.; Mo, S.Y.K. and Liu, A. (2015). “Twitter financial community sentiment and its predictive relationship to stock market movement”. *Quant. Finance, vol. 15, no. 10, pp. 1637–1656.*

Yuan, Y. (2015). “Market-wide attention, trading, and stock returns”. *Journal of Financial Economics 116(3): 548–564.*

Zhou, M.; Lei, L.; Wang, J.; Fan, W. and Wang, A. (2015). “Social media adoption and corporate disclosure”. *Journal of Information Systems, 2015.*

Zimbra, D.; Chen, H. and Lusch, R.F. (2015). “Stakeholder analyses of firm-related Web forums: Applications in stock return prediction”. *ACM Trans. Manag. Inf. Syst., vol. 6, no. 1, pp. 2:1 – 2:38.*

Zubair, S. and Cios, K.J. (2015). “Extracting news sentiment and establishing its relationship with the S&P 500 index”. *Hawaii International Conference on System Sciences, IEEE, 2015, pp. 969–975. doi: 10.1109/HICSS.2015.120*

2016

Dimpfl, T. and Jank, S. (2016). “Can internet search queries help to predict stock market volatility?”. *Eur. Financ. Manag., vol. 22, no. 2, pp. 171– 192.*

Ding, X.; Zhang, Y.; Liu, T. and Duan, J. (2016). “Knowledge-driven event embedding for stock prediction”. *Proceedings of 26th Int. Conf. Comput. Linguistics, 2016 , pp. 2133–2142.*

Eickhoff, M. and Muntermann, J. (2016). “Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom”. *Information & Management 53 (2016) 835–845.*

Feuerriegel, S. and Prendinger, H. (2016). “News-based trading strategies”. *Decision Support Systems 90 (2016) 65–74.*

Kharde, V. and Sonawane, S. (2016). “Sentiment analysis of Twitter Data: A Survey of Techniques”. *International Journal of Computer Applications, S. 5-15.*

Kumar, B.S. and Ravi, V. (2016). “A survey of the applications of text mining in financial domain”. *Knowl.-Based Syst., vol. 114, pp. 128–147.*

Li, Q.; Chen, Y.; Jiang, L.L.; Li, P. and Chen, H. (2016). “A tensor-based information framework for predicting the stock market”. *ACM Trans. Inf. Syst., vol. 34, no. 2, pp. 11:1– 11:30.*

Luo, B.; Zeng, J. and Duan, J. (2016). “Emotion space model for classifying opinions in stock message board”. *Expert Syst. Appl., vol. 44, pp. 138–146.*

Mäntylä, M.; Graziotin, D. and Kuutila, M. (2016). “The evolution of sentiment analysis: a review of research topics, venues, and top cited papers”. *arXiv preprint arXiv:1612.01556*

McLean, R.D. and Pontiff, J. (2016). “Does academic research destroy stock return predictability?” *Journal of Finance 71 (2016) 5–32*

Medovikov, I. (2016). “When does the stock market listen to economic news? New evidence from copulas and news wires”. *Journal of Bank. Finance, vol. 65, pp. 27–40*

Oliveira, N.; Cortez, P. and Areal, N. (2016). “Stock market sentiment lexicon acquisition using microblogging data and statistical measures”. *Decision Support Systems, vol. 85, pp. 62–73.*

Pröllochs, N.; Feuerriegel, S. and Neumann, D. (2016). “Negation scope detection in sentiment analysis: Decision support for news-driven trading”. *Decision Support Systems 88 (2016) 67–75.*

Shynkevich, Y.; McGinnity, T.; Coleman, S.A. and Belatreche, A. (2016). “Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning”. *Decision Support Systems, vol. 85, pp. 74– 83.*

Sinha, N.R. (2016). “Underreaction to news in the US stock market”. *Q. J. Finance, vol. 6, no. 2, pp. 1 650 005:1–1 650 005:46.*

Veiga, A.; Peiran J. and Ansgar, W. (2016). “Social Media, News Media and the Stock Market”. *News Media and the Stock Market (March 29, 2016).*

2017

Deng, S.; Sinha, A.P. and Zhao, H. (2017). “Adapting sentiment lexicons to domain-specific social media texts”. *Decision Support Systems 94 (2017) 65–76*

Ho, C.S.; Damien, P.; Gu, B. and Konana, P. (2017). “The time-varying nature of social media sentiments in modeling stock returns”. *Decision Support Systems 101 (2017) 69–81*

Janetzko, D.; Krauss, J.; Nann, S. and Schoder D. (2017). “Breakdown: Predictive Values of tweets, Forums and News in EUR/USD Trading.” *Proceedings of Thirty eighth International Conference on Information Systems (ICIS), Seoul.*

Kraus, M. and Feuerriegel, S. (2017). “Decision support from financial disclosures with deep neural networks and transfer learning”. *Decision Support Systems 104 (2017) 38–48*

Leitch, D. and Sherif, M. (2017). “Twitter mood, CEO succession announcements and stock returns”. *Journal of Computer Science 2017*

Li, T.; van Dalen, J. and van Rees, P.J. (2017). “More than just noise? Examining the information content of stock microblogs on financial markets”. *Journal of Inf. Technol. , pp. 1–20, 2017, https://doi.org/10.1057/s41265-016-0034-2*

Song, Q.; Liu, A. and Yang, S.Y. (2017). “Stock portfolio selection using learning-to-rank algorithms with news sentiment,” *Neurocomputing, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neucom*

Yang, S.Y.; Mo, S.Y.K.; Liu, A. and Kirilenko, A.A. (2017). “Genetic programming optimization for a sentiment feedback strength based trading strategy,” *Neurocomputing, vol. 264, pp. 29–41, 2017.*

2018

Chen, C.C.; Huang, H.H.; Shiue, Y.T. and Chen, H.H. (2018). “Numeral Understanding in Financial tweets for Fine-grained Crowd-based Forecasting”. *Accepted by the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2018), Santiago, Chile*

Feuerriegel, S. and Gordon, J. (2018). ”Long-term stock index forecasting based on text mining of regulatory disclosures“. *Decision Support Systems Volume 112, August 2018, Pages 88-97*

Li, Q.; Chen, Y.; Wang, J.; Chen, Y.Z. and Chen, H.C. (2018). “Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective". IEEE Transactions on Knowledge and Data Engineering, *vol. 30, no. 2, pp. 381-399.*

Mahmoudi, N.; Docherty, P. and Moscato, P. (2018). “Deep neural networks understand investors better”. *Decision Support Systems, 2018, Volume 112, Page 23*

Pröllochs, N. and Feuerriegel, S. (2018). “Investor Reaction to Financial Disclosures across Topics: An Application of Latent Dirichlet Allocation”. *Decision Sciences. doi:10.1111/deci.12346*

2019

Feuerriegel, S. and Gordon, J. (2019) „News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions.“ *European Journal of Operational Research 272(1):162-175, ISSN 0377-2217*

Contribution I

Krauss, J.; Nann, S.; Simon, D.; Fischbach, K. and Gloor, P.A. (2008) *"Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis"* was published in the proceedings of the European Conference on Information Systems (ECIS) as a completed research paper.

2008

Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis

Jonas Krauss
University of Cologne, jkrauss@smail.uni-koeln.d

Stefan Nann
University of Cologne, snann@smail.uni-koeln.de

Daniel Simon
University of Cologne, simond@smail.uni-koeln.de

Kai Fischbach
University of Cologne, kfischbach@wim.uni-koeln.de

Peter Gloor
Massachusetts Institute of Technology, pgloor@mit.edu

PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS THROUGH SENTIMENT AND SOCIAL NETWORK ANALYSIS

Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai; University of Cologne, Pohligstrasse 1, Cologne, Germany, {jkrauss,snann,simond}@smail.uni-koeln.de, kfischbach@wim.uni-koeln.de

Gloor, Peter, MIT, 3 Cambridge Center, Cambridge MA, USA, pgloor@mit.edu

Abstract

This paper introduces a new Web mining approach that combines social network analysis and automatic sentiment analysis. We show how weighting the forum posts of the contributors according to their network position allow us to predict trends and real world events in the movie business. To test our approach we conducted two experiments analyzing online forum discussions on the Internet movie database (IMDb) by examining the correlation of the social network structure with external metrics such as box office revenue and Oscar Awards. We find that discussion patterns on IMDb predict Academy Awards nominations and box office success. Two months before the Oscars were given we were able to correctly predict nine Oscar nominations. We also found that forum contributions correlated with box office success of 20 top grossing movies of 2006.

Keywords: Trend Prediction, Dynamic Social Network Analysis, Online Forum, Internet Movie Database, Oscar Awards

1 INTRODUCTION

It has been widely acknowledged that the "wisdom of crowds" as demonstrated in prediction markets (Surowiecki, 2004, Manski, 2006) is a surprisingly accurate mechanism to predict future trends. Large groups of "ordinary" people are better in predicting trends than a single expert. At the same time, the Web has turned into a major platform for information exchange, thus becoming a mirror of the real world: Millions of volunteers post latest news on Web sites such as Wikipedia, and political blogs such as dailykos and instapundit. In addition people express their opinions in forums and online communities, and tell openly what matters to them. Approaches such as "Netnography" (Kozinets, 2002) make use of this fact for marketing research, proposing analysis of statements of "devotees" and "insiders" in online forums and other Web sites. This paper proposes combining these two ideas, interpreting opinionated discussions and the level of "buzz" about the movie business on the Web as some kind of a prediction market.

Our approach offers an automated, efficient, and cheaper way to tap people's opinion than polling people over the phone. Our method calculates levels of "Web Buzz" by mining discussions in movierelated online forums, combining information about the structure of the social network with an analysis of the contents of the discussion. This paper demonstrates our approach by predicting the success of movies based on the communication in the online community IMDb.com. We analyze its communication patterns in regard to metrics like "intensity" and "positivity". While intensity means the frequency of the subject in discussion, positivity refers to the degree of positive feelings towards a movie expressed by contributors. Thereby we factor in quantitative and qualitative dimensions of discussion allowing us to extract an aggregated community opinion about individual movies.

These measurements are the basis of our two hypotheses. First, we assume that the chances of a movie to win an Oscar can be determined by the communication structure of the IMDb community. Second, we speculate that there is a relationship between the communication intensity about a movie and the performance of the movie at the box office.

The remainder of this paper is organized as follows. Section 2 surveys current research pertaining to movie success and the influence of word-of-mouth in online communities. Section 3 develops a

methodology to measure structural properties of online communities and to predict the success of a movie from these properties. Sections 4 and 5 apply our method to the online movie discussion forum IMDb.com.

2 RELATED WORK

There have been different approaches to examine the potential determinants of movie box office success. Most of the studies conclude that movie critics play a significant role for the success or failure of a film (Terry & Butler & De'Armond, 2005). Eliashberg and Shugan (1997) distinguish two possible perspectives on the role of critics: The influencer and the predictor perspective. From the first perspective critics are opinion leaders who influence their audience and, consequently, the box office performance of movies. The predictor perspective suggests that critics might be predictors of performance but not necessarily causing it. Dodds and Holbrook (1988) conducted an analysis where they compared influence and the effect of an Oscar nomination and movie critics on the success of a movie at the box office. Pardoe (2005) focused on models predicting nominees or winners at the Academy Awards.

Awad, Dellarocas and Zhang (2004) analyzed the influence of online movie ratings on box office success and developed statistical models based on these ratings to forecast movie revenues. Furthermore, they examined the relationship of traditional consumer communication, such as infomediary (professional critics), and online word-of-mouth versus offline word-of-mouth. They used the Internet Movie Database (IMDb, http://www.imdb.com) as their main source for the online data and determined the correlation between infomediaries and online word-of-mouth as well as infomediaries and offline word-of-mouth. Eventually, they came to the conclusion that online wordof-mouth has great potential for growth and an increasing number of consumers will use online rating and online review sites as the Internet becomes more pervasive. Surveying current critical issues in the motion picture industry, Eliashberg, Elberse and Leenders (2006) suggest further research relating Internet resources and movie consumption as well as box office sales.

Research regarding trendsetters (Clark & Zboja & Goldsmith, 2007, Valente, 1996) is often associated with the concept of social network analysis (Wasserman & Faust, 1994). One prominent concept is the one of information cascades (Bikhchandani & Hirshleifer & Welch, 1992, Anderson & Holt, 1996, Anderson & Holt, 1997, Bikhchandani & Hirshleifer & Welch, 1998) which explains convergent behaviour patterns and therefore holds potential to identify trends and trendsetters. However, other experiments showed only limited validity of the concept being applied to different laboratory setups (Huck & Oechssler, 2000, Hung & Plott, 2001). Trendsetters have also been of great interest for quite some time in the field of marketing where Myers and Robertson (1972) discuss the importance of "opinion leadership". Connected to opinion leaders is the concept of social contagion which describes the spreading of behavior patterns in a community (Burt, 1987, Crandall, 1988, Rodgers & Rowe, 1993, Kretschmer & Klimis & Choi, 1999). Yet, contagion of opinion does not necessarily result from social influencers, also marketing actions can induce the spread of a certain opinion (Bulte & Lilien, 2001). While IMDb has been frequently used as a basis to predict movie success by other researchers (Eliashberg & Sawhney, 1996, Jensen & Neville, 2002, Pardoe, 2005, Simonoff & Sparrow, 2000, Dellarocas & Awad & Zhang, 2007, Kaplan, 2006), little research has been done so far in using communication behavior and social network structure of an online community as a determinant of movie success at the box office and as a predictor for Oscar nominations.

Although Awad, Dellarocas and Zhang (2004) base their model on movie ratings of an online community, they do not make use of further information which could be retrieved through an analysis of the patterns of communication in that community. Our approach enhances prior research by taking into account social network structures in an online community and by measuring discussion content rather than movie ratings.

3 OUR APPROACH

Predicting real world events based on the communication structure and contents of online word-ofmouth networks is a rapidly emerging field (Patak et. al, 2007, Ganiz & Pottenger & Yang, 2007). This paper

contributes to this field by using methods of social network analysis and web data mining to run a model for forecasting movie success.

We chose two Web data sources for our research; namely IMDb, and the "Box Office Mojo" (www.boxofficemojo.com) webpage. In our analysis we focused on the message board community of IMDb. This community exclusively discusses movie and theater related topics and has over 4 million users (Big Boards, 2007) making it the biggest online movie community. With at least 15 million monthly unique U.S. visitors in 2007 (Compete, 2008), IMDb considerably outperforms other online movie communities in terms of its traffic. Additionally, amongst the biggest online movie communities IMDb is the only one having a dedicated subforum for discussion of topics related to the Academy Awards. As mentioned above, IMDb – with the message board community being an integral part – has been subject of extensive research in the past and has also gained wide recognition in public, e.g. by being labelled as one of the "25 Sites We Can't Live Without" in 2007 (Time Magazine, 2007).To measure box office success, movie release dates and movie show times we used Box Office Mojo. In our work this information was compared to the IMDb message board communication structure. A social link between two participants in a forum is constructed if an answer to a message is also an answer to all messages previously posted. Our analysis is based on all posts in a forum from December 2005 to December 2006.

Based on the communication retrieved from the IMDb message boards we applied a model consisting of three relevant components: Discussion intensity, positivity and time. While intensity and time seem to be easy to analyze, the degree of positivity expressed in the discussion requires a more sophisticated approach. Various authors used the degree of positive emotions expressed in communication for deriving insights about the topic of their analysis (Bales, 1950, Bales & Cohen & Williamson, 1979, Gottman & Rose & Mettetal, 1982, Echeverria, 1994, Losada & Fredrickson, 2005). Their results show that discussion positivity can be a key factor for analysis. We will describe our application of intensity, positivity and time in the next section.

The research and application of the model was done with the Condor social network analysis tool, formerly called TeCFlow (Gloor et al., 2003). Condor creates visual maps, movies and many graph metrics of relationships related to social networks by mining web site link structures, online forums and e-mail networks. For example, Condor can create graphical static link views of the communication between users in a web forum and calculates the actor contribution index (Gloor et al., 2003), which delivers clues about the relevance and importance of key actors contributing to the communication. For this paper we make use of Condor's two main features: Firstly, it allows analyzing continuous temporal changes in communication structures in a web forum. Secondly, it supports content analysis of terms being used in forum communication, which also can be displayed graphically in a static or dynamic view.

For further comparison of our results we used the online version of the Linguistic Inquiry and Word Count (LIWC, www.liwc.net) software which offers features to rate textual inputs according to their emotional properties.

4 ACADEMY AWARDS FORECAST BASED ON COMMUNITY COMMUNICATION

The goal of our first experiment was to pick likely candidates for the Oscar Academy Awards, given end of February 2007, based on an analysis of the forums on the IMDb concluded end of December 2006.

As our first hypothesis suggests, we assume that a correlation between the Academy Awards presentation for a particular movie and the communication about that movie in the IMDb forums exists. We speculate that communication intensity and quality of the discussion about a particular movie are indicators of a movie being nominated for an Academy Award. While it would be very hard to also predict in what category a movie would receive an award, we will show that we are able to predict if a movie will be a candidate for an award.

As the basis of our analysis we used the "Oscar Buzz" forum, which is a subforum of IMDb. In this forum topics related to the Academy Awards are being discussed by the IMDb community. This forum has a high frequency of readers and message posters (500 to 1000 posts per day).

To analyze the communication in the Oscar Buzz forum we ran a series of Condor queries, with data from November and December 2006. From the resulting list of terms we extracted the top 25 movies that were discussed in the subforum. We then counted the number of times they were mentioned as well as the time span from their release date to December 15th, 2006.

We based our computation of the chance of a movie being nominated for an Academy Award on three factors. The three factors consist of two temporal frequency indices, the "Intensity Index" and the "Positivity Index" as well as a temporal noise factor, the "Time Noise Factor".

The Intensity Index measures the degree of communication intensity about a specific topic. It is calculated for each movie separately. The Intensity Index is a normalization of the "numbers of mentions" on a scale of 0 to 1. The index is calculated by dividing each value by the highest value of the compared movies (table 1). By identifying this index we followed the approach of Frank & Antweiler (2004) who found a significant correlation between the amount of messages being posted about stocks in finance-related online forums and their volatility. Although this study deals with a different subject, there are similarities in terms of the underlying technology and the communication patterns of online communities. Therefore, we assume that the more a movie is talked about in the community the higher is its chance to receive a nomination for an Oscar. This fact is acknowledged by our model by comprising the numbers of mentions in form of the Intensity Index as a component of the model.

The second index measures the quality of the communication about a certain movie, in particular how positive the communication about this movie was. To calculate the Positivity Index we used the content processing function of Condor for finding out if the discussion about the movie was associated with positive terms. These terms have been determined by ranking potential phrases in regards to their betweenness centrality with Condor's content processing function. The highest ranked terms then became our actual positivity phrases: "win," "nominate," "great," "good," "award," "super," "oscar," and "academy". We selected those terms because they show that the discussion about a particular movie is carried out under positive aspects regarding its Oscar nomination and they are the most important positive phrases in aspects of betweenness centrality. Our method follows the "bag-ofwords" concept, which basically means that the order of words in a document can be neglected (Aldous, 1985). This approach makes no direct use of grammatical structure. In previous research it has been found that only a small increase in accuracy is gained by attempting to exploit grammatical structure in the algorithms (Frank & Antweiler, 2004). However, there are cases where this approach might lead to a wrong result: If a negation of a positive term is used in a forum post (e.g. "not a good movie") our method will still give it a positive rating. In the future we plan to further adapt the sentiment analysis algorithm in order to exclude these cases; however in this project our results show that even this simple approach leads to a good prediction quality.

This approach is similar (to a degree) to the one which is used by the developers of the software LIWC who determine the positivity of a text through comparing it with a dictionary (LIWC, 2007). When comparing our positivity index with LIWC using a random sample of IMDb posts, we found significant correlation between LIWC and positivity index ($R=0.56$, $p<0.01$).

Intensity and Positivity Index are not fully independent: the number of positive terms mentioned in context with a movie will increase with the number of messages about this movie. However, it is also possible that a movie will be talked up in a negative context. To prove this we would also need to incorporate a "Negativity Index". This will be a necessary extension for further research.

Movie	Intensity Index	Positivity Index	Time Noise Factor	Oscar Model
Apocalypto	0,05	0,15	0,02	0,15
Babel	0,46	0,30	0,16	0,30
Blood Diamond	0,24	0,24	0,02	0,24
Bobby	0,19	0,20	0,07	0,20
Borat	0,24	0,24	0,13	0,24
Departed	1,00	1,00	0,22	1,00
Dreamgirls	0,52	0,45	0,00	0,45
Flag of our Fathers	0,29	0,20	0,18	0,20
James Bond: Casino Royale	0,21	0,18	0,09	0,18
Little Children	0,61	0,37	0,22	0,37
Little Miss Sunshine	0,50	0,28	0,45	0,28
Open Season	0,35	0,23	0,24	0,23
Pirates of the Caribbean	0,17	0,14	0,51	0,14
Pursuit of Happiness	0,16	0,13	0,00	0,13
Stranger than Fiction	0,31	0,15	0,11	0,15
Take the Lead	0,39	0,51	0,80	0,51
Thank you for Smoking	0,14	0,15	0,87	0,15
The Break Up	0,20	0,15	0,62	0,15
The Devil wears Prada	0,16	0,13	0,53	0,13
The Nativity Story	0,23	0,15	0,04	0,15
The Prestige	0,11	0,15	0,18	0,15
The Queen	0,59	0,41	0,24	0,41
United 93	0,55	0,24	0,73	0,24
V for Vendetta	0,09	0,14	0,87	0,14
When a Stranger calls	0,25	0,15	1,00	0,15

Table 1. Factor values of top 25 movies.

An interesting insight of our positivity analysis using Condor is that the terms “oscar”, “win” and “nomin” always build a ring structure in the communication about the movie. This means that these three terms are mostly mentioned together.

For the computation of the Positivity Index each Positivity Term was given a relevance value for its influence on the discussion. As mentioned above, three terms are strongly linked and always built a ring. Reflecting the “term frequency inverse document frequency” weight (tfidf), this means that those three terms share a great amount of posts and are therefore of great significance (Salton & Buckley, 1988, Gloor & Zhao, 2006). This is why we chose the highest values for those terms and gave lower values to the remaining terms. “Frequency” consists of the number of times a term was associated with a movie. The Positivity Index in table 1 is computed by the following formula:

$$\textit{Positivity Index} = \sum_{\textit{positivity terms}} \textit{relevancevalue} * \textit{frequency}$$

The resulting Positivity Indices are then normalized on a scale from 0 to 1, which leads to the values in the column “Positivity Index” as listed in table 1. For calculating the Positivity Index we used the weights of the term-to-term relationships that factor in the betweenness centrality (Wasserman & Faust, 1994) values of the related terms. Thus, the weights do not just correspond to the frequency of terms. Through the implicit application of the graph drawing algorithm of Fruchterman and Reingold (1991), which is implemented in Condor, also the “importance” of the terms is measured. This algorithm is used to construct the social network and calculate centrality values of the participating actors, in this case the corresponding terms of the positivity network.

The last of the three factors we used for determining the Oscar Model is a noise factor that takes into account that some movies are older than others. This models the fact that discussion of a movie calms down over time in the message boards. Nelson, Donihue, Waldman and Wheaton (2001) also find strong evidence regarding the industry practice of delaying movie releases until late in the year as it improves the chances of receiving nominations. Therefore we introduced a “Time Noise Factor” to our model. It is being calculated by normalizing the days from the movie release date till December 15th, 2006 on a scale from 0 to 1 for all of the 25 movies. December 15th, 2006 is the date where the latest of the 25 movies being subject to our investigation was released. The values of the Time Noise Factor can be looked up in table 1.

To determine the Oscar Model, our predictor for the probability of a movie getting an Academy Award nomination, each of the previously calculated indices, Intensity , Positivity and Time Noise is weighed by a factor:

$$Oscar\ Model = a * + b * + c * + \mid a + b + c = 1$$

We empirically determined the best values for these factors by running all possible factor combinations (with steps of 0.1) against the known Oscar outcome. The results suggest that setting b to 1 and a and c to 0 leads to an optimal solution. Figure 1 shows the plotted curves for the different factor combinations. When applying the Oscar Model to a real world event we included an error term S. By looking up the actual Oscar winners and nominees for the movies of all factor combinations we minimized S, what can be expressed by the number of movies that neither received an Oscar nor a nomination. In the optimal setting five out of the top ten movies ranked by the Oscar Model received an Oscar and four received a nomination for an Oscar (table 1).

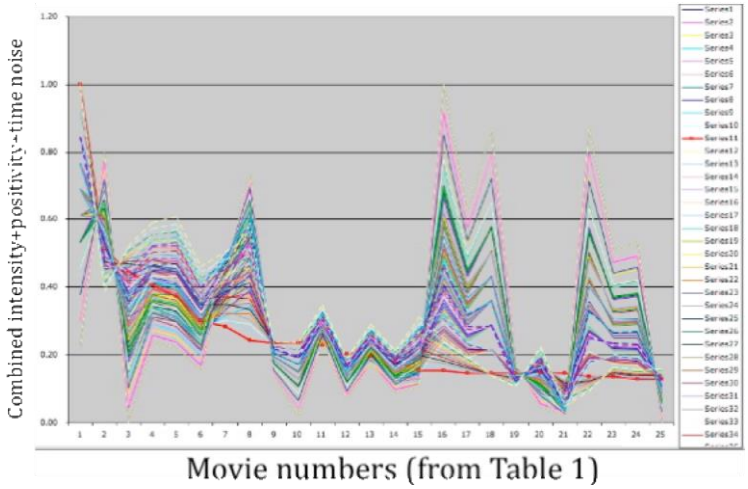

Figure 1. Oscar Model Sensitivity Analysis.

Weighing b with 1 delivered the best result with 9 out of the top 10 movies ranked by the Oscar Model being actual award winners or nominees respectively (red line in figure 1, series 11; for Award winners and Oscar Model values refer to table 2). Interestingly, the best factor combination is therefore the one ignoring intensity and time noise. This comes from the types of users participating in the discussion on IMDb, whom we suspect to be movie buffs and therefore more in line with the opinion of the Academy Awards jury than others.

This shows that movies that are being discussed in a positive way in the sub forum “Oscar Buzz” have a high probability of getting a nomination for the Academy Awards. It further indicates that the users who are participating in the communication in “Oscar Buzz” are movie enthusiasts who value similar criteria in a movie as the Oscar poll does. As shown by an Oscar Index twice as high as the next movie, there is a clear favorite for the Oscar nomination in the IMDb community, namely “The Departed”. The community opinion (reflected by the values of the Oscar Model) is not limited to only a few movies but rather a broad range of movies is being discussed intensively (table 1, Intensity Index). As stated earlier there is indeed a correlation of 0,88 between the intensity of discussion and the Positivity Index , yet it is the positivity index which is the best predictor of winning an Oscar. Moreover, due to the Time Noise Factor being included in our computation, an aggregation of the indices in a multiplicative model does not appear to be applicable. In the worst case such a multiplicative model could lead to highly positive discussed movies receiving an Oscar Model value of 0 if being released on December 15th, 2006. Thereby the Time Noise Factor would be significantly overvalued.

Top 10 Oscar Model	Model Value	Actual Result	LIWC Value
Departed	1,00	Oscar	3,85
Take the Lead	0,51	-	6,58
Dreamgirls	0,45	Oscar	7,57
The Queen	0,41	Oscar	6,96
Little Children	0,37	Nomination	3,44
Babel	0,30	Oscar	6,67
Little Miss Sunshine	0,28	Oscar	3,58
United 93	0,24	Nomination	6,26
Borat	0,24	Nomination	5,32
Blood Diamond	0,24	Nomination	2,82

Table 2. Values of the Oscar Model Vs. Academy Award results.

In order to compare our results with other available methodologies for analyzing the positivity of communication, we repeated the same analysis with the above mentioned LIWC software. However, we found no correlation (R=0.065, non-significant) between the results computed by LIWC and the values of the Oscar Model. A possible explanation might be that LIWC uses a general dictionary as opposed to our customized method of calculating the Positivity Index. Table 2 lists the values of LIWC.

It should be pointed out that there are different categories of Oscar Awards. There are six major ones that people primarily focus on: best picture, best director, and the four acting awards (best actor/actress, best supporting actor/actress). Hard core film buffs may also talk about second-tier awards like best screenplay or editing or music, and the other awards in the more technical arts (Art Direction, Sound, etc.), but these are not typically the subjects of most of the buzz. What we found is that the importance of the awards movies got corresponds to the level of buzz. "Babel" with a lower value for the Oscar Model won an award for best score, which is a minor Oscar. By contrast "Departed" with the highest value won two major awards (Best Picture and Best Director) and also two important second tier awards (Best Editing and Best Adapted screenplay). "Little Miss Sunshine," which won for best actor and best original screenplay, the first a slightly more prominent award than the second, but still not in the same rank as best picture and best director, also has a value for the Oscar Model slightly higher than "Babel," but much lower than "The Departed."

The results of this first application of our approach encouraged us to apply the same model to the prediction of a movie's box office success. We will describe the procedure of adjusting the model for this application in the next section.

5 CORRELATION BETWEEN MOVIE SUCCESS AND COMMUNITY COMMUNICATION

Based on our findings that intensive and positive online forum communications are predictors for Oscar success, we applied the same insights to predict commercial success of not yet launched movies. To study movie success at the box office we chose the IMDb sub forum "Previews & Reviews". As our metrics of financial success we analyzed the US movie box office rankings of 2006, which we obtained from Box Office Mojo. The major success criterion of a movie we used in this analysis is its gross sales at the box office in 2006. We concentrated on twenty films, which prevailed in the community discussion in the "Previews & Reviews" IMDb forum and also showed top ranks in the 2006 gross sales list.

Movie	Intensity Index	Positivity Index	Trendsetter Index	Values of the Buzz Model	Box office success in $
Pirates of the Caribbean: Dead Man's Chest	1,00	1,00	1,00	1,00	423.315.812
Cars	0,62	0,67	0,88	0,68	244.082.982
Superman Returns	0,76	0,67	1,00	0,78	200.081.192
Ice Age: The Meltdown	0,29	0,67	0,75	0,50	195.330.621
Casino Royale	0,49	1,00	1,00	0,75	167.220.102
Over the Hedge	0,51	0,33	1,00	0,55	155.019.340
The Departed	0,95	0,67	1,00	0,87	132.208.177
Borat	0,25	0,17	0,88	0,35	128.488.700
Dreamgirls	0,18	0,33	0,88	0,37	102.266.997
Inside Man	0,56	0,17	0,88	0,51	88.513.495
Monster House	0,33	0,33	0,75	0,41	73.661.010
Underworld: Evolution	0,27	0,50	0,50	0,39	62.318.875
Little Miss Sunshine	0,62	0,83	0,88	0,73	59.863.257
Blood Diamond	0,22	0,33	0,88	0,38	56.576.961
The Queen	0,27	0,00	0,88	0,31	54.581.202
The Prestige	0,29	0,33	0,88	0,42	53.089.891
Apocalypto	0,27	0,50	1,00	0,49	50.866.635
Stranger than Fiction	0,25	0,50	0,88	0,45	40.435.190
Snakes on a Plane	0,29	0,00	0,63	0,27	34.020.814
Friends with Money	0,29	0,17	0,25	0,25	13.368.437

Table 3. Indices and box office gross sales of top 2006 movies.

Our goal was to develop an appropriate metric to measure the communication behavior of the community regarding movies. Therefore, using Condor's content processing capabilities and following our general approach of analyzing communication in regards to intensity and expressed positivity, we created three individual indices that capture the communication patterns of the users in the subforum. We again used Intensity Index and a Positivity Index. A new metric was introduced with the Trendsetter Index, all three indices were combined into a "Buzz Model".

To calculate the Trendsetter Index we first identified users with the highest betweenness centrality values in the sub forum "Previews & Reviews". With a minimum value of 0.03, the betweenness centrality of these 10 identified users was at least 12 times higher than the average betweenness centrality of 0.0025. In a second step we counted for each movie how many trendsetting users were mentioning the movie favorably. The index was then calculated by normalizing the number of participating trendsetters on a scale of 0 to 1, which is in line with the calculation of the Intensity Index. This metric implicitly emphasizes the social aspects of the communication. It weighs the impact of the most between users in the conversation and is an indicator of the importance and influence of trendsetters on the communication about a certain movie in the forum. We speculate that discussion of these trendsetters will likely have a direct impact on the success of a movie at the box office. Table 3 displays all three indices and the values calculated with Condor.

We used a similar formula as for the Oscar Model to determine a combined "Buzz Model" with Intensity, Positivity, Trendsetter, and Error Term S:

$$BuzzModel = a * + b * + c * + \mid a + b + c = 1$$

To determine the optimal values for a, b and c we ran all possible factor combinations (in steps of 0.1) against 20 top grossing movies in 2006. At values a = 0,5, b = 0,3 and c = 0,2 correlation is 0.75 ($p<0.01$), showing a very strong relationship between the communication intensity/behavior and the box office success of movies. Despite a positive correlation with the Intensity Index (R=0.44) and the Positivity Index (R=0.42), the Trendsetter Index does not become superfluous and obviously contributes to the optimal solution.

While analyzing IMDb.com it was obvious that certain movies were significantly more discussed than others. The question was if there would be a relationship with the financial success of the movie or if the discussion at imdb.com would be independent from the "real" world.

In our analysis we found robust support for our hypothesis that higher movie success correlates with higher communication intensity. A positive discussion about a movie in the forum correlates with higher

revenue of the movie at the box office. This means that high positive discussion by trendsetters predicts success of a newly released movie at the box office.

Furthermore, we have seen that the most influential (high-betweenness) users lead the discussion, which indicates that this discussion may have an impact on the result of the movie at the box office. More in-depth analysis shows, however, that this opposite conclusion can not be proven. Our analysis does not tell if "talking up" a movie will guarantee financial success. The IMDb.com community consists of movie experts who are not showing the same attitude towards a movie as the average moviegoer. The value of the Buzz Model of the movie "Snakes on a Plane" illustrates this point. This movie was "hyped up" long before its release throughout the web, yet in the discussion on IMDb.com it received comparably bad press, which shows that IMDb.com users are clearly more differentiated in their perception of the movie than the mainstream user was, and more resistant to attempts of manipulation by movie publishers.

Note that this paper is not focusing on the general discussion of the effects of "Buzz" in a community. This might be subject of a more in-depth analysis of online communities and part of a continuation of this article.

6 LIMITATIONS

Our research is subject to limitations, which, though they do not affect the positive results in this paper, need to be tackled through further adaption of the model. One aspect, which has been mentioned already, is the "bag-of-words" concept. This needs to be resolved through the application of a context sensitive method, which takes into account the actual relation between phrases in the analyzed communication. The quality of the sentiment analysis could be further increased through broader sensitivity analysis of the potential phrases.

The results of last year's Academy Award could be predicted relatively well (though the award category was not predicted). However, results should be scrutinized by applying the same model to future Oscar elections. With respect to the Buzz Model, results could be re-evaluated by applying a multiplicative model. Another point left to discuss is the causality chain: Is movie success determined by forum discussion or does forum communication follow movie success? Our approach only calculates the correlation between these two, yet the underlying reason for the correlation remains unclear.

7 SUMMARY AND CONCLUSION

This paper represents an extension of the research on the influence of online communities on the success of movies. It is addressing two main issues: First it introduces a model to predict Academy Award nominees based on the communication of an online community. It then applies the same approach to examining if there is a correlation between community communication and movie success at the box office. Doing so, we were able to make predictions about real world events based on social networks in an online movie community.

In our first experiment we showed that there is a correlation between the IMDb community discussion and the chance of a movie getting nominated for an Academy Award. Some insights could be gained about the structure and properties of the community in the Oscar Buzz sub forum of IMDb. Oscar influencers are movie buffs who do not necessarily have the same opinions as mainstream movie viewers. With "The Departed" a clear favorite of the forum for getting a nomination for an Oscar was identified 8 weeks before the Oscars were awarded.

In our second experiment we found that a high intensity of discussion about a particular movie at IMDb is a strong indicator of success of that movie at the box office. While not every movie being successful at the box office is actively discussed in the community, every movie, which generates high positive buzz on IMDb appears high in the box office charts. This means that high discussion volume predicts success at the box office, but generating lots of buzz will not help a movie to increase viewing in theaters. For Oscars, just gauging the level of positivity of posts is enough to predict future success. Using a customized dictionary yields better results than a generic positivity measurement tool such as LIWC.

For predicting financial success, on the other hand, a more complex model assigning higher weight to trendsetters, i.e. people with central network positions, delivers the best results.

The insights we gathered and the methods we apply could be of value also in the field of marketing science, especially in the field of viral marketing. For example, motion picture studios could optimize their marketing strategy through identifying trendsetters in forums and the internet and then address those with their marketing campaigns. Forum communication analyzed by our methodology could be used as an indicator for early success prediction of an upcoming movie release. These few examples show the practical relevance of our analysis, ideas of connected research are suggested below.

Our experiments can be extended in different ways in future research. An obvious extension would be to increase the sample size by widening the data analysis over longer periods of time and by including other forums. It would be of great interest whether including other forums would entail an even higher correlation or whether those forums would perform worse in terms of predictive qualities, thereby strengthening our perception of IMDb being an expert community. Secondly, it would be interesting to examine whether similar insights could be obtained for other movie genres as well. For example, one could focus on the discussion about TV shows and compare the communication structure in the forums with audience ratings. These approaches could be easily used as an indicator in the movie business to predict which movies, TV shows, etc. would be successful in the future. Thus, IMDb message boards and similar forums could be used as a market research platform for all kinds of movierelated predictions. It might be interesting to apply our approach of quantifying unstructured communication to motion picture business external fields using blogs or newsgroups and trying to make predictions about other real world events based on communication taking place in these groups.

For example, it would be of great interest to apply our Oscar prediction model to other award nominations to test the model with other data sets. We are also currently applying the same model to online investor forums to predict financial performance of selected stocks. Although our approach worked well predicting this year's Academy Awards and movie box office success, it will need much further work to get a more robust proof of its predictive qualities.

8 ACKNOWLEDGEMENTS

We are grateful to Rob Laubacher for helping us understand the dynamics in the movie business and educating us on the differences between „major" and „minor" Oscars. We also thank Ken Riopelle for pointing us to the LIWC sentiment analysis software.

9 REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII — 1983, 1–198, Springer, Berlin.

Anderson, L. R. and Holt, C. A. (1996). Classroom Games: Information Cascades. Journal of Economic Perspectives, 10, 187-193.

Anderson, L. R. and Holt, C. A. (1997). Information Cascades in the Laboratory. The American Economic Review, 87 (5), 847-862.

Awad, N. F. and Dellarocas C. and Zhang X.(2004). Is Online Word-of-mouth a Complement or Substitute to Traditional Means of Consumer Conversion. Sixteenth Annual Workshop on Information Systems Economics (WISE), Washington, DC.

Bales, R. F. (1950). Interaction Process Analysis: A Method for the Study of Small Groups. AddisonWesley.

Bales, R. F. and Cohen, S. P. and Williamson, S. A. (1979). SYMLOG: A System for the Multiple Level Observation of Groups. Free Press.

Big Boards (2007). IMDb statistics. http://www.big-boards.com/board/926, retrieved 2007.

Bikhchandani, S. and Hirshleifer, D. and Welch, I. (1992). A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades. Journal of Political Economy, 100, 992-1026.

Bikhchandani, S. and Hirshleifer, D. and Welch, I. (1998). Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades. The Journal of Economic Perspectives, 12 (3), 151170.

Bulte, C. and Lilien, G. (2001). Medical Innovation Revisited: Social Contagion versus Marketing Effort. The American Journal of Sociology, 106 (5), 1409-1435.

Burt, R. S. (1987). Social Contagion and Innovation: Cohesion Versus Structural Equivalence. The American Journal of Sociology, 92 (6), 1287-1335.

Clark, R. A. and Zboja, J. J. and Goldsmith, R. E. (2007). Status consumption and role-relaxed consumption: A tale of two retail consumers. Journal of Retailing and Consumer Services, 14 (1), 45-59.

Compete (2008), SnapShot of imdb.com. http://siteanalytics.compete.com/imdb.com/?metric=uv, retrieved 2008.

Crandall, C. S. (1988). Social Contagion of Binge Eating. Journal of Personality and Social Psychology. 55 (4), 588-598.

Dellarocas, C. and Awad, N. F. and Zhang, X. (2007). Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures. Working Paper, Robert H. Smith School Research Paper.

Dellarocas, C. and Narayan, R. A. (2005). Statistical Measure of a Population's Propensity to Engage in Post-purchase Online Word-of-mouth. R. H. Smith School of Business, University of Maryland, College Park, MD 20742, Working Paper.

Dodds, J. C. and Holbrook, M. B. (1988). What's an Oscar worth? An Empirical Estimation of the Effect of Nominations and Awards on Movie Distribution and Revenues. Current Research in Film: Audiences, Economics and the Law, 4.

Echeverria, R. (1994). La Ontologia de1 Lenguaje. Dolmen Ediciones, Santiago de Chile.

Eliashberg, J. and Elberse, A. and Leenders, M. (2006). The Motion Picture Industry. Marketing Science, 25 (6), 638-661.

Eliashberg, J. and Sawhney, M. S. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. Marketing Science, 15 (2), 113-131.

Eliashberg, J. and Shugan, S. M. (1997). Film critics: Influencers or Predictors?. Journal of Marketing, 61 (2), 68-78.

Frank, Murray Z. and Antweiler, Werner (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance, 59 (3), 1259-1294.

Fruchterman, T. M. J. and Reingold, E. (1991). Graph Drawing by Force-Directed Placement. Software-Practice and Experience, 21 (11),1129-1164.

Ganiz. M. and Pottenger, W. M. and Yang, X. (2007). Link Analysis of Higher-Order Paths in Supervised Learning Datasets. In Proc. 5th Workshop on Link Analysis, Counterterrorism and Security, SIAM International Data Mining Conference.

Gloor, P. A. and Laubacher, R. and Dynes, S. B. C. and Zhao, Y. (2003). Visualization of Communication Patterns in Collaborative Innovation Networks: Analysis of some W3C working groups. In Proceedings of the Twelfth International Conference on Information and Knowledge Management.

Gloor, P.A. and Zhao, Y. (2006). Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis. In Proceedings of 10th IEEE International Conference on Information Visualisation IV06.

Gottman, J. M and Rose, F. and Mettetal, G. (1982). Time-series analysis of social interaction data. Emotion and Early Interaction, 261-289.

Huck, S. and Oechssler, J. (2000). Informational cascades in the laboratory: Do they occur for the right reasons?. Journal of Economic Psychology, 21, 661-671.

Hung, A. A. and Plott, C. R. (2001). Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions. The American Economic Review, 91 (5), 1508-1520.

Jensen D. and Neville J. (2002). Data mining in social networks. Invited presentation to the National Academy of Sciences Workshop on Dynamic Social Network Modeling and Analysis, p. 7-9, Washington, DC.

Kaplan, D. (2006). And the Oscar Goes to… A Logistic Regression Model for Predicting Academy Award Results. Journal of Applied Economics & Policy, 25 (1), 23-41.

Kozinets, R. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. Journal of Marketing Research, 39, 61-72.

Kretschmer, M. and Klimis, G. M. and Choi, C. J. (1999). Increasing Returns and Social Contagion in Cultural Industries. British Journal of Management, 10 (1), 61–72.

Linguistic Inquiry and Word Count (2007). The LIWC2001 Application. http://www.liwc.net/liwcdescription.php, retrieved 2007.

Losada, M. and Heaphy, E. (2004). The role of positivity and connectivity in the performance of business teams: A nonlinear dynamics model. American Behavioral Scientist, 47 (6), 740–765.

Manski, C. F. (2006). Interpreting the Predictions of Prediction Markets. Economic Letters, 91, 425429.

Myers, J. H. and Robertson, T. S. (1972). Dimensions of Opinion Leadership. Journal of Marketing Research, 9, 41-46.

Nelson R. A. and Donihue M. R. and Waldman D. M. and Wheaton C. (2001). What's an Oscar Worth?. Economic Inquiry, 39 (1).

Pardoe, I. (2005). Predicting Academy Award winners using discrete choice modeling. In Proceedings of the 2005 Joint Statistical Meetings, Alexandria, VA. American Statistical Association.

Patak, N. and Mane, S. and Srivastava , J. and Contractor, N. (2007). Knowledge Perception Analysis in a Social Network. In Proc. 5th Workshop on Link Analysis, Counterterrorism and Security, SIAM International Data Mining Conference.

Rodgers, J. L. and Rowe, D. C. (1993). Social contagion and adolescent sexual behavior: A developmental EMOSA model. Psychological Review, 100 (3), 479-510.

Salton, G. and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval (1988). Information Processing and Management, 24 (5), 513-523.

Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. Chance 13 (3), 15-24.

Surowiecki, J. (2004). The Wisdom of Crowds. Doubleday, New York.

Terry, N. and Butler M. and De'Armond D. (2005). The Determinants of Domestic Box Office Performance in the Motion Picture Industry. Southwestern Economic Review, 32 (1), 137-148.

Time Magazine (2007), 25 Sites We Can't Live Without. http://www.time.com/time/specials/2007/article/0,28804,1638266_1638253_1638236,00.html, retrieved 2008.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18 (1), 69-89.

Wasserman, S. and Faust, K. (1994). Social Network Analysis, Methods and Applications. Cambridge University Press.

Wolfers, J. and Zitzewitz, E. (2004). Prediction Markets. Journal of Economic Perspectives, 18 (2)

Contribution II

Gloor, P.A.; Krauss, J.; Nann, S.; Fischbach, K. and Schoder, D. (2009) *"Web Science 2.0: Identifying trends through semantic social network analysis"* was published in the proceedings of the International Conference on Computational Science and Engineering (IEEE) as a completed research paper

2009 International Conference on Computational Science and Engineering

Web Science 2.0: Identifying Trends through Semantic Social Network Analysis

Peter A. Gloor
Center for Collective Intelligence
MIT, Cambridge, MA, USA
pgloor@mit.edu

Jonas Krauss, Stefan Nann
University of Applied Sciences
Northwest Switzerland Brugg
{jonas.krauss, stefan.nann}
@fhnw.ch

Kai Fischbach, Detlef Schoder
Dept. of Information Systems
University of Cologne
{fischbach, schoder}@wim.uni-koeln.de

Abstract—**We introduce a novel set of social network analysis based algorithms for mining the Web, blogs, and online forums to identify trends and find the people launching these new trends. These algorithms have been implemented in Condor, a software system for predictive search and analysis of the Web and especially social networks. Algorithms include the temporal computation of network centrality measures, the visualization of social networks as Cybermaps, a semantic process of mining and analyzing large amounts of text based on social network analysis, and sentiment analysis and information filtering methods. The temporal calculation of betweenness of concepts permits to extract and predict long-term trends on the popularity of relevant concepts such as brands, movies, and politicians. We illustrate our approach by qualitatively comparing Web buzz and our Web betweenness for the 2008 US presidential elections, as well as correlating the Web buzz index with share prices.**

Social network analysis, semantic social network analysis, trend prediction, Web mining

I. INTRODUCTION

The Internet has become a major communication channel for late-breaking news and to disclose innermost secrets. For example, when CBS published documents about George W. Bush's behavior during his military service, Republican bloggers quickly identified weak spots in the authenticity of the documents. This questionable evidence regarding George Bush's potential evasion of military service during the Vietnam War era ultimately lead to the early retirement of CBS news anchor Dan Rather. This incident is just one of many illustrating that today's news are made and disseminated on the Web and in the blogosphere. The Web therefore has become both part of and a mirror of the "real world". Assuming that people will be doing what they announce, analyzing what influential people say on the Web might identify trends before they have been recognized by the rest of the world [15]. Towards this goal, we introduce a new way of measuring the changes in popularity of brand names and famous people such as movie stars, politicians, and business executives, based upon the premise that in today's Internet economy, buzz on the Web reflects popularity and buzz in the real world.

The approach described in this paper mines and analyzes unstructured communication and information from Web resources. As input for our method we take concepts in the form of representative phrases from a particular domain – for example names of politicians, brands, or issues of general interest. In a first step the geodesic distribution of the concept in its communication network is determined by calculating the temporal betweenness centrality of the linking structure. The second step adds the social network position of the concept's originator – called "actor" in social network language – to the metric to include context-specific properties of nodes in the social network. In the third step we qualitatively evaluate the concept's communication context to assess the concept's perception on the Web, blog, or online forum.

Result of this three-step process is a "Web buzz index" for a specific concept that allows for an outlook on how the popularity of the concept might develop in the future. In the remainder of this paper, after an overview of the state of the art, we introduce our three-step process. We illustrate it first by tracking the presidential elections, and then by showing the correlation between fluctuations in the Web buzz index for stock titles and stock prices.

II. RELATED WORK

Popularized by Barabási [4] in his book "Linked", there is a rich body of research on how the linking structure of the Web influences accessibility of Web pages and their ranking in search engines.

Visualization of Web structure and contents has been an active area of research since the creation of the Web. There are numerous systems for the static visualization and analysis of the link structure of the Web [9], [10]. Inxight, Visual Insight, Touchgraph, Grokster, and Mooter are all systems for the visualization of the linking structure of the Web, sometimes also offering a visual front end for search results.

In a related stream of work, researchers have been trying to predict the hidden linking structure based on known links [1], [2]. Additionally, by looking at contents of Web sites, subspaces of the Web have been clustered by topics [6]. Combining these two lines of research, community Web sites have been mined to discover trends and trendsetters for viral marketing [18].

Our research focuses on a similar application – tracking the strengths of concepts over time. For our analysis we are using the Condor system [13] originally developed to mine email networks to automatically generate dynamic social network movies.

There are various studies that are dealing with the prognosis of stock prices through an analysis of online

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.186

communication in blogs and message boards [22], [3]. Researchers are also basing their studies on the most popular finance-related online communities Yahoo! Finance, Raging Bull, and Motley Fool [14]. References [8], [21], [22] are applying sentiment extraction algorithms on finance-related communication data from message boards [24].

III. Categorization of Web Sources in Information Spheres

For our Web mining approach we classify the World Wide Web into three categories or information spheres: The Web at large – we call it "Wisdom of Crowds", the blogosphere – "Wisdom of Experts", and forums – "Wisdom of Swarms". Each of these three sources is processed differently in our method based on the way how the information contained in it is produced. Online forums contain the most focused and upto-date information about a certain subject. These forums are self-organized communities consisting of individuals as well as organizational institutions, which exchange ideas and information [20]. The huge "swarms" of people in the forum represent the collective opinion of those who care most about the forum's topic.

Blogs represent the "Wisdom of Experts". The number of bloggers and new blogs grew exponentially over the last few years and is still growing. Contrarily to forums, where posters engage in a dialogue amongst themselves, bloggers are individual experts where each of them is expressing his or her private opinion. Because an expert is not always right, it would be risky to rely on a single opinion. But combining the wisdom of experts about a subject will lead to an aggregated indicator of the collective opinion of experts about a certain topic.

Finally, mining the Web at large also gives valuable clues about a certain topic. The topics might be discussed on sites of varying popularity and actuality such as online news sites, company Websites, information Websites, etc. This resource is by far the largest of the three and incorporates the collective opinion of a large part of the Western world – what we call the
"Wisdom of the Crowds."

These three different data sources represent the basis for our combined communication and information analysis process.

IV. CONCEPT WEIGHTING STRATEGY

For the last six years we have developed a sophisticated semantic social network analysis tool called Condor [12], [13]. Condor (formerly called "TeCFlow") includes automated textual analysis functionality using standard information retrieval algorithms like "term frequency–inverse document frequency" [19]. Additionally, Condor factors in the betweenness centrality of actors for weighing the content by the social network position of actors.

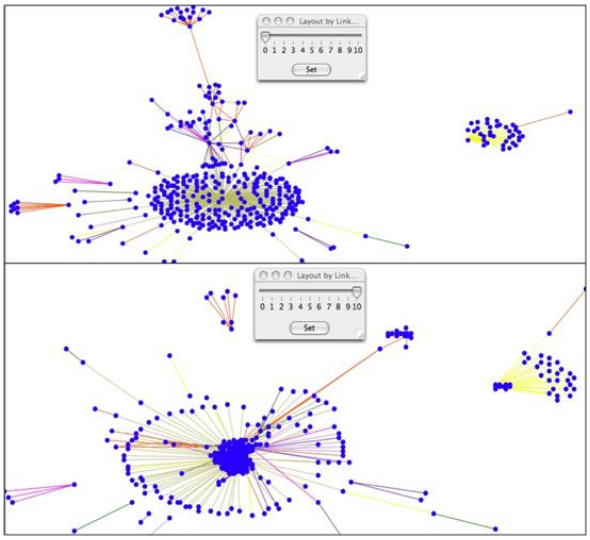

Figure 1. Weighting a set of documents by social network position of actors only (top), and also factoring in similarity of contents. All networks in this and subsequent figures are visualized with the Fruchterman-Rheingold graph layout algorithm [11].

Fig. 1 illustrates this concept by showing two Condor screen shots of the same document network. The top of the picture shows a social network of actors based on exchange of e-mails. While senders and receivers of e-mails are represented by nodes, the edges reflect an exchange of e-mails between two actors. The bottom of the picture shows the same network, but now the actors have additionally been grouped by the similarity of contents of their discussion. The blue and very dense cluster in the middle of the network represents all actors that are talking about the same subject in their e-mail communication. Clustering of nodes at the bottom of fig. 1 is therefore done by combining two attractive forces, first based on the number of exchanged e-mails, and second based on the similarity between two e-mail text bodies calculated by "term frequency–inverse document frequency".

Thus both shared vocabulary of the social network and actors' network position are factored in in the results of the textual analysis of Condor. In the next three sections we will describe our three-step approach: "What – Who – How". "What" stands for the concepts we are extracting and measuring over time. The "Who" represents the actors using the concepts we want to track, while the "How" measures the positive or negative sentiment in which the actors use the concepts. Determining the social network position of actors – the "Who" – has different semantics for each of the three information spheres. On the Web, the betweenness of actors is measured by the linking structure of the Web pages pointing back to pages talking about them. In the blogosphere, we only consider links to other blog posts as a measure of confidence of other bloggers into the poster of the original blog post. In online forums, the relative importance of a poster is based on the communication structure and the poster's position in the social network.

V. What - Measuring Temporal Betweenness of Concepts

The first step to measuring a trend is the tracking of a concept's relative importance in a relevant information sphere – Web, blog, or online forums. As an approximation for the relative importance of a concept in the information sphere, we calculate the betweenness centrality of this concept within the chosen information sphere. This means that we are extending the well-known concept of betweenness centrality of actors in social networks to semantic networks of concepts.

Betweenness centrality of a concept in a social network is an approximation of its influence on the discussion in general. Betweenness centrality in social network analysis tracks the number of geodesic paths through the entire network, which pass through the concept whose influence is measured. As access to knowledge and information flow are means to gain and hold on to power, the betweenness centrality of a concept within its semantic networkis a direct indicator of its influence [23]. In other words, concepts of high betweenness centrality are acting as gatekeepers between different domains. While communication in online forums can be used to construct social networks among actors, we can also construct social networks from blogs and the Web. Although these semantic networks based on blog and Web links are not true social networks in the original sense, they are straightforward to construct by considering theWebsites and blog posts as nodes and the links between the Websites and blog posts as ties of the social network.

Measuring the betweenness centrality of a concept permits us to track the importance of a concept in the chosen information sphere. This can be done either as a one-time measurement, or continuously in regular intervals over time, as Web pages, blog posts, and forum posts all have time stamps. We therefore periodically (e.g. once per day, once per hour, etc.) calculate the betweenness centrality of the concept. The resulting betweenness centrality is a numerical value between zero and one, with zero implying no importance of the concept in the information sphere and values above zero representing the relative importance in comparison to other concepts.

To build the semantic social network in an information sphere we introduce degree-of-separation search. Degree-ofseparation search works by building a two-mode network map displaying the linking structure of a list of Web sites or blog posts returned in response to a search query, or the links among posters responding to an original post in an online forum. For example, a search to get the betweenness of "Hillary Clinton" on the Web works as follows:

1) Start by entering the search string "Hillary Clinton" into a search engine.
2) Take the top N (N is a small number, for example 10), of Web sites returned to query "Hillary Clinton".
3) Get the top N Web sites pointing to each of the returned Web sites in step 2 by executing a "link:URL" query, where URL is one of the top N Web sites returned in step 2. The "link:" query returns what the search engine considers "significant" Web sites linking back to a specific URL.
4) Get the top N Web sites pointing to each of the returned Web sites in step 3. Repeat step 4 up to the desired degree of separation from the original top N Web sites collected in step 2. Usually it is sufficient, however, to run step 4 just once.

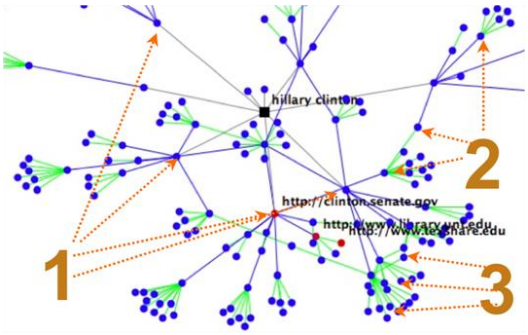

Figure 2. Degree-of-separation search for "Hillary Clinton"

Fig. 2 illustrates the two-mode network map returned to the query "Hillary Clinton". The level-0 node is the query term, level-1 nodes are the URLs connected directly to the query, i.e. the original search results. Level-2 nodes are the most highly ranked search results returned by the "link" query, to each of the top N level-1 nodes. Level-3 nodes are the most highly ranked nodes returned by the "link" queries of each of the level-2 nodes.Fig. 2 gives a visual overview of the betweenness of each of the level-1 and level-2 nodes. The more links a node has pointing to it, the more between it is. For example the node labeled http://clinton.senate.gov is linked by a group of level 2 nodes which themselves are linked by groups of level-3 nodes. This indicates that the node http://clinton.senate.gov will have fairly high betweenness itself.

Fig. 3 illustrates how degree-of-separation search can be used to compare the relative importance of the concepts "gun control", "abortion", "gay marriage", and "Iraq war". This means the importance of an individual concept depends on the linking structure of the temporal network and the betweenness of the other concepts in the network. Condor queries for each concept were run on the Web in 2006, when the war in Iraq was dominating US headlines. Fig.3 shows the semantic social network combining the search results for these four concepts.
A degree-of-separation search for several concepts always results in a fully connected graph since Websites such as Wikipedia or New York Times connect all resources. This is because usually among the level-1 nodes, but at the latest among the level-2 nodes, there will be Wikipedia and other top-rankend Web sites, acting as connectors.

Betweenness values for each concept are calculated in the connected graph formed by combining the Web links pointing to the top ten search results for each of the four Web queries by running a degree-of-separation search for each of the four search queries.

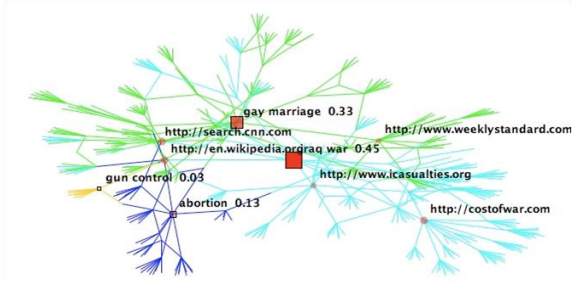

Figure 3. Comparison of the importance on the Web of „gay marriage", "gun control", "abortion", and "Iraq war". Squares are query terms, circles are URLs, size denotes betweenness

The war in Iraq dominates the discussion, followed by gay marriage. Gun control was almost a non-issue at that time,with a centrality factor less than a tenth of the war in Iraq.We can also see that costofwar.com, www.weeklystandard.com, Wikipedia, and cnn.com are the Web sites with the highest betweenness centrality.This also explains why we get a fully connected graph when we combine the four networks for the four concepts: there are always very central, i.e. highly between Web sites such as Wikipedia connecting seemingly unrelated concepts,thus permitting us to calculate betweenness for each concept in comparison to the others.

Note that this ranking has nothing to do with the absolute number of search hits returned by the search engine. If a concept has been around for a long time, it will have accumulated many Web pages, therefore leading to many hits. A newly emerging "hot" concept, which appears on highranked Web sites, will not necessarily have that many hits, but will have high betweenness.

Measuring trends is not restricted to measuring popularity of abstract concepts, but can easily be applied to measuring popularity of people. The next example illustrates the "Web popularity" among the top seven Republican and seven Democratic contenders to become the next US President, as of end of August 2006. Fig. 4 shows the combined degree-ofseparation search results for 10 US Presidential hopefuls. Each of the colors identifies the set of nodes and links between them retrieved from the information sphere for one of the presidential candidates, e.g. the Web sites and links returned to concept "Al Gore" are shown in blue. While the red squares represent the search queries the red nodes are the Web sites returned by more than one query. The bigger a node the more important it is in the relative network. The relative position of two concepts inside the network to each other can be interpreted as "how close in substance" two concepts, i.e. two presidential candidates are to each other.

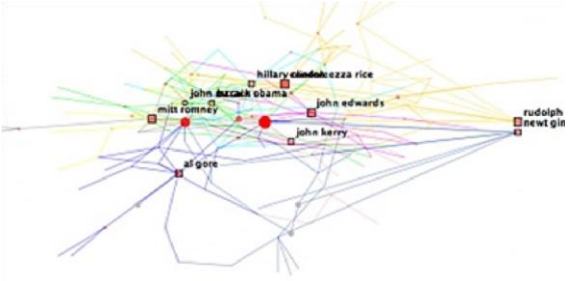

Figure 4. Degree-of-separation searches combined for presidential hopefuls in Aug 2006

For example, in fig. 4, Rudolph Giuliani and Newt Gingrich seem to go off together "to the far right". Table I lists the results of the two most recent presidential polls as of end of August 2006 and compares them with the betweenness values of the candidates on the Web calculated in September 2006.

TABLE I. POLLS AND RELATIVE WEB BETWEENNESS FOR US PRESIDENTIAL CANDIDATES IN 2006

Democrats	**Pew Aug 9-13**	**Am.Polling June 13-16**	**Betweenness Web Aug 26**
Hillary Clinton	40%	36%	0.05
Al Gore	18%	-	0.10
John Edwards	11%	15%	0.10
John Kerry	11%	13%	0.05
Joseph Biden	6%	4%	0.02
Bill Richardson	4%	5%	0.06
Russ Feingold	2%	6%	0.01
Republicans			
Rudolph Giuliani	24%	21%	0.09
Condoleezza Rice	21%	30%	0.04
John McCain	20%	20%	0.03
Newt Gingrich	9%	8%	0.05
Mitt Romney	4%	7%	0.02
George Allen	-	5%	0.03
Bill Frist	3%	2%	0.06

Based on the poll values in table I, we would expect Hillary Clinton and Rudy Giuliani to be the most between actors in our Web analysis. The result is slightly different, however. While there are no surprises for Rudy Giuliani, Hillary is not really the top ranked democratic candidate by betweenness. This honor falls to Al Gore and John Edwards, who are tied for first place. The reason for non-candidate Al Gore's surprising popularity were the recent launch of his new movie "An Inconvenient Truth" about global warming, generating buzz for Al Gore not only as a politician, but also as a movie actor and environmentalist. Al Gore therefore connects different Web communities, or in the language of social networks, he bridges structural holes, leading to high betweenness. Al Gore's high betweenness also illustrates that comparing relative betweenness only makes sense among similar concepts – such as US Presidential candidates in our example.

Repeating the calculations periodically over time permits to measure changes in betweenness of the different candidates to identify trends. This temporal concept importance is the foundation for steps two – "Who" and three – "How" of our approach.

Fig. 5 illustrates the changing betweenness values of the 14 presidential contenders over 14 days. As the blue line shows, non-competing candidate Al Gore's lead is growing, while other leading democratic candidate John Edward's fortunes are declining. The big winner of the first week is candidate Russ Feingold, whose absolute betweenness and thus Web popularity is more than doubling before going down again in the second week. Leading republican candidate Rudy Giuliani is keeping his lead, in a neck-on-neck race with Al Gore. The overall centrality of the combined group analysis is slightly diminishing over the

time period, indicating that there is no clear leader emerging thus far.

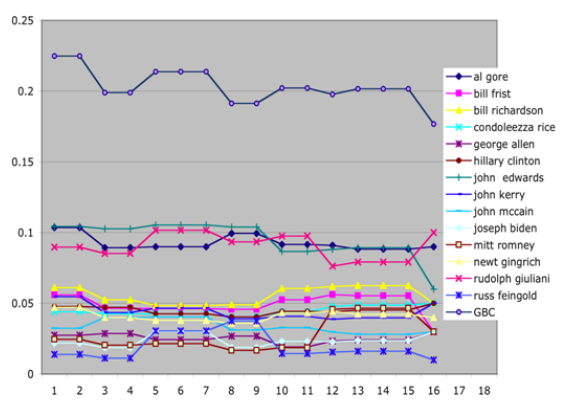

Figure 5. Web buzz trend over 18 days in August 2006 of US Presidential candidates

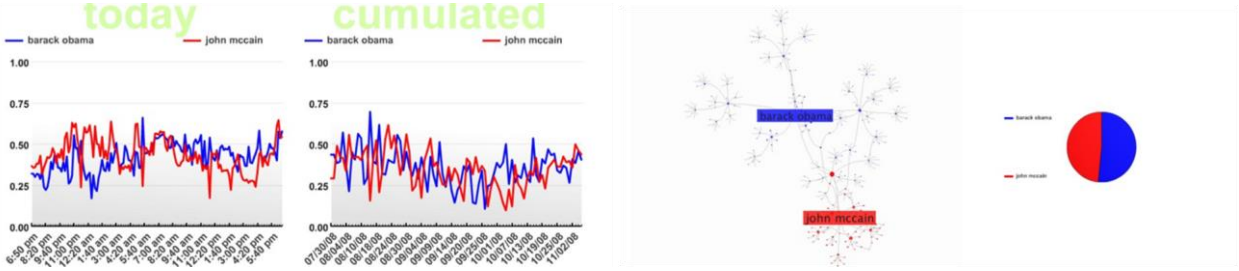

Figure 6. Blog buzz trend with Condor right after November 4, 2008 of US Presidential elections

Fig. 6 illustrates the Blog buzz right after the US presidential elections Nov 4, 2008. Democrat Barack Obama won the elections against Republican John McCain with a landslide in electoral votes (365 against McCain's 162) and 53% of the popular vote. Fig. 6 illustrates this process measuring the betweenness of search strings "John McCain" and "Barack Obama" in the blogosphere. The upper left window shows the minute-by-minute readings, which, at the time of the election, change by the minute based on new posts about either of the candidates on high betweenness blogs such as the Huffingtonpost or Powerlineblog. The overall trend favoring Barack Obama, the blue line, can however clearly be seen. In the accumulated graph, in the upper right window, starting in September, Obama's betweenness line consistently trumps over McCain's betweenness. The bottom left picture shows the social network of blog posts. The blogs talking about McCain form a far more compact cluster, at the very bottom with a tightly interlinked structure. The democratic blogs, linking to Obama, are much wider spread out, and also exhibit fewer interconnecting links, reflecting the wider political interests of the voters supporting Obama. The pie chart at the lower right shows the relative betweenness of the two candidates, 53% for Obama, against 47% for McCain). Note that these relative betweenness numbers correspond to the percentages for the candidates in the popular vote.

VI. Who - Weighing Discussion Content by the Social Network Position of Actors

The "Who" step is based on the idea that what certain people say carries more weight, i.e. that some people are more influential than others. As an approximation of their influence we use their betweenness. In the "Who"-step of our approach we add a context-specific weight of the concept's importance, based on the importance of the actor, which is using the phrase. Depending on the information sphere, the actor is either a Web site, a blog (standing in for the respective blogger), and the poster in the online forum. Thereby we factor in that not all actors in the network are equal and that their importance matters for the discussion of the concept.

The context-specific importance of a phrase is based on its originator's betweenness centrality. By multiplying the betweenness centrality of the actor with the betweenness centrality of the concept we factor in the influence of the actor in the information sphere. This not only supports the elimination of spam that might have been produced to "game the system", but also introduces "expert ranking".

Looking at the presidential candidates example (Fig. 4, 5 & 6) the evaluation of the semantic social network by betweennesspermits to find the most relevant Websites. These Websites can be regarded as "kingmakers" in our context. Kingmakers are Web sites that, through linking to a concept, increase the betweenness of the original concept through their own high betweenness centrality. In our presidential polling analysis, en.wikipedia.org and www.ovaloffice2008.com are the most between Websites. While it is not surprising that Wikipedia is very central, as all candidates take care to get their profiles entered and updated there, the central position of ovaloffice2008 comes as somewhat of a surprise. For each individual network generated by the degree-of-separation search for each candidate, Wikipedia, the candidates' own Websites, and the sites of national newspapers such as the New York Times or the Washington Post rank higher. If the Websites returned to the 14 different degree-of-separation queries are combined, however, a different picture emerges, with Wikipedia and ovaloffice2008 by far having the highest centralities. While the Google page rank of Wikipedia is 9 (out of 10), ovaloffice2008's Google page rank [5] in August 2006 was only 5. Its betweenness in the context of presidential elections, however, is the second highest of all Websites included in this analysis of presidential hopefuls. Ovaloffice2008 also includes a very active forum where citizens of different inclinations and party colors discuss strengths, weaknesses, and chances of the various candidates, motivating the central position of this Website. Note that we will always get one connected network when combining the individual Web networks, as there are the "superconnectors"

like Wikipedia and New York Times, linking the individual networks of the candidates.

VII. How - Determine Discussion Quality through Sentiment Analysis

Measuring temporal betweenness of concepts in online communication and weighing the content with the importance of actors provides new possibilities of identifying and analyzing new trends. However, there is a third component that needs to be incorporated into the process. It is not only about What and Who, it is equally important to look at positive and negative emotions in the discussion. For our final example we loaded communication data for 21 stock titles from the finance-related online community Yahoo! Finance into Condor. Yahoo! Finance offers individual message boards for several thousand companies from various industries. The way actors are talking about a particular subject can be determined through sentiment analysis. Besides visualizing and measuring temporal betweenness of concepts or actors Condor includes effective text analysis methods. Through its content process functionality the software automatically can identify most frequent words and word pairs in large amount of texts. [7] has shown that automatic extraction of words and word pairs leads to more precise results than manually selecting positive and negative words. We have implemented a two-step approach first using the automatic term extraction algorithm of Condor to get most relevant words and word pairs. In the second step we created term lists of words and word pairs with positive and negative sentiment by reading the extracted bags of words. The lists are concept dependent and were specifically selected for the analyzed company. Condor provides the possibility of applying stop word lists to exclude common words like "the", "for" or "and". After the identification of positivity and negativity lists we then extracted the frequency of company related, positive, and negative terms within posts. The combination of these three metrics – frequency, positivity, and negativity – represents the sentiment of the forum users on a company. The following table shows the term lists for Goldman Sachs. To enhance the significance and accuracy of the text analysis we implemented an algorithm based on regular expressions. The algorithm detects and analyzes co-occurrence of company terms with positive or negative words in a forum post. This makes it possible to identify the sentiment about a subject in a forum on a particular day.

TABLE II. Company terms, positivity and negativity lists for Goldman Sachs

Company terms	Positivity list	Negativity list
gs, goldmansachs, goldman, sachs	better, bought, buy, buy puts, buy shares, buy stock, buy stocks, buying, earnings, going higher, good, good time, higher prices, investment, long, longs, profits, won	back, bad, didn, dont, down, going down, inflation, little, losses, lower, market down, recession, sell, selling, short, short position, shorting, shorts, sold, stock down

The approach follows the basic "bag-of-words" approach which is also considering co-occurrence of keywords in sentences or text [17]. A drawback of this approach is the disregard of grammatical dependencies in the analyzed data. This might lead to misleading interpretation in some cases. For example the statement "Goldman is not good" would be classified as a positive sentiment with the simple "bag-ofwords" approach. In practice this problem seems to be rare, however. Reference [16] states that 40% of analyzed keywords in the same sentence or text block show grammatical dependencies. By reading a large sample of forum messages we empirically verified their finding that actors mostly use negative phrases rather than negating positive phrases when they want to express something negative. For example they use the phrase "is bad" instead of "is not good".

VIII. Combining the "What-Who-How": the Web Buzz Index

To test our approach combining social network data from all three information spheres, we collected data over 213 days (April, 1st 2008 until October, 30th 2008) on 21 stock titles on Yahoo! Finance. Additionally, we tracked the temporal Web and blog betweenness for the same titles with Condor. We implemented an algorithm that determines correlation between the Web buzz and actual stock price. The Web buzz is comprised of Web and blog betweenness, and forum sentiment. Forum sentiment is calculated through the metrics introduced in section 7: term frequency, positivity, and negativity. Each of these metrics has been calculated in two different ways: the simple way only considering sentiment and a second way weighing the sentiment with the social network position of an actor. This makes it possible to weigh forum posts by the "importance" of the poster. This classification results in eight indices, Web betweenness, Blog betweenness, Positivity, Positivity betweenness, Negativity, Negativity betweenness, Wordcount (representing frequency), and Wordcount betweenness.

We calculated index values for a time window of 30 days. We smoothened the index curves by moving averages from five to twelve days. The results for Goldman Sachs are shown in fig 7.

We observed that on days where the stock price rose the negativity indices were inversely correlated. The same was true for the positivity indices on days where the stock price fell. This means that on days with rising stock price the inverse of the negativity index has to be taken, while on days with falling stock price the inverse of the positivity index was taken. As fig. 7 illustrates at time window size of 30 days average correlation values are significant at levels of 0.05 ($n = 30$, $r > 0.361$). We found that the highest value for the moving average most often showed optimal results. Generally, the correlation values of the indices in time window 30 show that a relation between Web buzz and stock price movement exists.

Fig. 8 shows the individual plotted curves cumulatively making up the Web buzz index in relation to the stock price for a moving average of 12 days.

Time Window: 30 days									
	Moving Average								
	5	6	7	8	9	10	11	12	max
Index Type	Average correlation of sub time periods with stock prices								
Wordcount	0.350	0.369*	0.384*	0.392*	0.392*	0.388*	0.388*	0.394*	0.394*
Wordcount Betweenness	0.421*	0.433*	0.440*	0.442*	0.442*	0.441*	0.440*	0.448*	0.448*
Positivity	0.334	0.351	0.361*	0.366*	0.371*	0.374*	0.380*	0.386*	0.386*
Positivity Betweenness	0.409*	0.416*	0.421*	0.424*	0.425*	0.424*	0.424*	0.430*	0.430*
Negativity	0.331	0.344	0.357	0.361*	0.366*	0.368*	0.372*	0.380*	0.380*
Negativity Betweenness	0.406*	0.412*	0.417*	0.417*	0.417*	0.416*	0.415*	0.420*	0.420*
Web Betweenness	0.321	0.322	0.317	0.313	0.314	0.331	0.342*	0.350	0.350
Blog Betweenness	0.348	0.370*	0.383*	0.394*	0.403*	0.414*	0.424*	0.427*	0.427*
(α=0.05)*									

Figure 7. Correlations between stock price and the different components of the Web Buzz Index for Goldman Sachs

IX. Outlook and Conclusion

In this paper we have shown that buzz on the Web mirrors the real world. Tracking concepts on the Web by differentiating between the Web at large, blogs, and online forums, and combining what people say with their social network position indeed permits to discover trends, frequently before the real world has become aware of them.

There remain some issues that deserve further investigation. The first concerns our dependence on the rankings of the search results by the search engine. We have used Google, Google Blog Search, MSN Search, and Yahoo.

While the top n Web sites about a topic returned by the different search engines vary, we found surprising consistency in the relative betweeness values of the search topics. We explain this through the presence of central Web sites such as Wikipedia, Yahoo, and the New York Times Web site in the resulting link networks. These Web sites always come up in the searches at one or two degrees of separation to the search topic, providing a consistent linking structure. The second issue is about causality. While we have demonstrated correlation between Web buzz and real-world events and have demonstrated the predictive capabilities of our approach for political elections and Oscars [19], more work needs to be done to formally show causality for stocks.

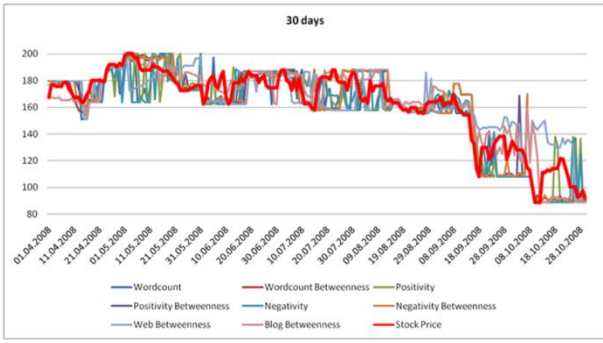

Figure 8. The 8 Web Buzz Indices plotted against the stock price of Goldman Sachs

We are currently testing our system in different application areas, trying to increase the accuracy of our political predictions and stock trend correlations. Possible extensions of our approach are the addition of the concept of fading in and out of new ideas. Frequently, new ideas are brought up by visionary people, only to lay dormant for extended periods of time until they are finally picked up by larger groups of people. We speculate that extending our model to incorporate this process might increase the correlation between Web buzz and the real world events we are trying to track. We also intend to explore whether different groups of actors based on their network position have different influence on Web communication. This would ultimately also lead to more accurate predictions. We are currently improving our sentiment extraction methods with additional algorithms, e.g. dynamically enhancing positivity and negativity lists with machine learning techniques. To further increase the predictive quality, we consider approaches such as applying a dynamic time offset between the Web buzz index and the stock price, and including industry indices and trading volumes.

Another idea is to combine the Web buzz analysis with prediction markets [25], by setting up automated agents trading in prediction markets based on Web buzz analysis. Extending this line of research, human participants in prediction markets could be given access to our Web trend prediction results in order to increase the quality of the prediction market.

Our vision is to develop a general system for trend prediction, identifying new ideas early on while they are being raised by the trendsetters. At this stage, new ideas have not yet been recognized by the rest of the world, but discovering them can be extremely valuable. Applications of our system might be for politicians trying to find out what the real concerns of their constituency are, or for financial regulators trying to identify micro- and macro-trends in financial markets.

ACKNOWLEDGMENT

We would like to thank Hauke Führes for patiently adding all our change wishes to Condor for the Web Buzz Index. We are indebted to Manfred Vogel for providing us with an excellent IT infrastructure at University of Applied Sciences North West Switzerland in Brugg, and to Tom Malone for insightful discussions on the properties of gatekeeper words in semantic networks.

REFERENCES

[1] Adar, E., Zhang, L., Adamic, L., Lukose, R. (2004) Implicit Structure and the Dynamics of Blogspace: Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference.

[2] Al Hasan, M., Chaoji, V., Salem, S., & Mohammed Z. (2006) Link Prediction using Supervised Learning: Proc 2006 Workshop on Link Analysis, Counterterrorism and Security.

[3] Antweiler, W., Frank, M. (2004) "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance. Vol. LIX, No. 3.

[4] Barabasi, L. (2003) Linked: How Everything Is Connected to Everything Else and What It Means. Plume.

[5] Brin, S. Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine, In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia: Elsevier.

[6] Chakrabarti, S., Joshi, M., Kunal, P., Pennok, D. (2002) The Structure of Broad Topics on the Web.Proc, WWW 2002, Hawaii.

[7] Das, S. R., Chen, M. Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, Management Science Vol. 53 Issue 9: p 1375-1388.

[8] De Choudhury, M., Sundaram, H., John, A., Seligmann, D. (2008) Can Blog Communication Dynamics be correlated with Stock Market Activity? Hypertext 2008, Pittsburg: PA – 19 to 21 [9] Dodge, M., Kitchin, R. (2002) Atlas of Cyberspace, Pearson Education.

[10] Dodge, M. Kitchin, R. (2000) Mapping Cyberspace, Routledge.

[11] Fruchterman, T. M. J., Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11).

[12] Gloor, P., Zhao, Y., (2006) Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis, Proceedings of 10th IEEE International Conference on Information Visualisation IV06 (London, UK, 5-7 July 2006)

[13] Gloor, P., Zhao, Y. (2004) TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis, ACM CSCW Workshop on Social Networks (ACM CSCW Conference, Chicago, 6. Nov. 2004).

[14] Jones, A. L. (2006) Have internet message boards changed market behavior?, info, Vol. 8 No. 5 2006, pp. 67-76.

[15] Kleinberg, J. (2008) The Convergence of Social and Technological Networks, Communications of the ACM, Vol. 51 No. 11 November 2008, pp. 66-72.

[16] Matsuzawa, H.; Fukuda, T. (2000). "Mining Structured Association Patterns from Databases," Proceedings of the 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining (2000), pp. 233-244.

[17] Nasukawa, T., Morohashi, M., Nagano, T. (1999) Customer claim mining: Discovering knowledge in vast amounts of textual data. Technical report, IBM Research, Japan, 1999.

[18] Richardson, M., Domingos, P. (2002) Mining Knowledge Sharing Sites for Viral Marketing, Proc. ACM SIGKDD, 2002.

[19] Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5): 513–523.

[20] Tapscott, D.; Williams, A. D. (2006), Wikinomics: How Mass Collaboration Changes Everything, Portfolio Hardcover, New York, 2006.

[21] Tetlock, P. (2007) Giving content to investor sentiment: The role of media in the stock market, Journal of Finance, Forthcoming.

[22] Tumarkin, R., Whitelaw, R. F. (2001) “News or Noise?“: Internet Message Board Activity on Stock Prices, Financial Analysts Journal, 57: pp. 41-51.

[23] Wasserman, S., Faust, K. (1994) Social Network Analysis, Cambridge University Press.

[24] Wysocki, P. D. (1999) “Cheap Talk on the Web: Determinants of Posting on Stock Message Boards”. University of Michigan: Working Paper, November 1999.

[25] Zitzewitz, E. Wolfers, J. (2004) “Prediction Markets”, Journal of Economic Perspectives, Winter 2004

Contribution III

Nann, S.; Krauss, J.; Schober, M.; Gloor, P.A.; Fischbach, K. and Führes, H. (2009) *"Comparing the structure of virtual entrepreneur networks with business effectiveness"* was published in the proceedings of the conference on Collaborative Innovation Networks (COINS) as a completed research paper.

Major work for this article was done during a research program at MIT. The article is also officially listed as a MIT Sloan School of Management working paper under the title *"The Power of Alumni Networks - Success of Startup Companies Correlates With Online Social Network Structure of Its Founders"*.

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2009) 000–000

Procedia

Social and Behavioral Sciences

www.elsevier.com/locate/procedia

Collaborative Innovation Networks (COINS) 2009

Comparing the structure of virtual entrepreneur networks with business effectiveness

Stefan Nann[a], Jonas Krauss[a], Michael Schober[b], Peter A. Gloor[a,*], Kai Fischbach[a], Hauke Führes[b]

[a]*MIT Sloan School of Management, 5 Cambridge Center, Cambridge, MA, 02142*
[b]*University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany*

Elsevier use only: Received date here; revised date here; accepted date here

Abstract

In this paper we look at the effectiveness of business networks created by alumni of different universities. In particular, we analyze the networking behavior of entrepreneurs in Germany through the emergent structures of their virtual social networks. We automatically collected the publicly accessible portion of the German business networking site Xing.com by crawling the Web. We then filtered the people by attributes indicative of their university, and roles as founders, entrepreneurs, and CEOs. We constructed alumni networks of 12 German universities, identifying over 50,000 alumni, out of which more than 15,000 had entrepreneurship attributes. We also manually evaluated the financial success of a subsample of 80 entrepreneurs for each university.

Universities, which are more central in the German university network provide a better environment for students to found more and more successful startups. People in alumni networks whose members have a stronger "old-boys-network", i.e. a larger share of their links with other alumni of their alma mater than with outside people, are more successful as founders of startups. We repeated this analysis on the individual level, combining all 15,000 founders, confirming the same result. Finally, the absolute amount of networking matters, i.e. the more links entrepreneurs have, and the higher their betweenness in the online network of university alumni, the more successful they are.

Keywords: online social networks, founder networks, startup success

1. Introduction

In this paper we look at the social networks forming around alumni of universities. In particular, we analyze the networking behavior of entrepreneurs in Germany through the emergent structures of their online social networks. Most of these founders are part of the generation of the twenty to forty year olds who are making heavy use of the

* Corresponding author. Tel.: +1-617-253-7018
E-mail address: pgloor@mit.edu.

Internet. According to the Pew Internet Survey (Jones & Fox 2009) over half of the adult Internet population in the US is between 18 and 44 years old, and using the Internet for entertainment and social networking. Likewise, a study by "Forschungsgruppe Wahlen" (2009) reveals that 72 percent of the adult German population uses the Internet (over 90 percent of the people between 18 and 49). These studies show that Blogs, Facebook, MySpace, LinkedIn, and Twitter have become major means of communication to stay in touch with friends and business partners, complementing established communication channels such as e-Mail and the phone. While private interaction on social networking platforms has become an active field of research (Ellison et. al. 2007; boyd 2008, boyd & Ellison 2008), little research had been done on the commercial value of keeping business contacts on social networking platforms such as LinkedIn (O"Murchu et al. 2004).

In this project we look at the entrepreneurial success of alumni of 12 major German universities. We analyze the relationship network of entrepreneurs as it is represented in the German social networking site Xing, investigating if social networking structure predicts entrepreneurial success.

This article is structured as follows. Section 2 presents the related literature and illustrates the reasons for extending this stream of research. In the same section we develop four research hypotheses. Section 3 describes the data collection and method employed. Section 4 highlights the findings. Finally, sections 5-7 discuss the theoretical and managerial implications of the findings, note their limitations, and provide some suggestions for further research.

2. Related Work and Hypotheses

We try to answer the research question if certain types of online social networking structures of entrepreneurs predict an entrepreneur"s success. Based on prior work on comparing social networking structure of individuals and companies with successful outcome of their work activities we would indeed expect that such a correlation exists. Research on this topic has investigated the effect of network structures on the performance of the individual (e.g. Ahuja et al. 2003; Bulkley and Van Alstyne 2006; Cross and Cummings 2004; Gloor et al. 2008; Mehra et al. 2001; Moran 2005; Sparrowe et al. 2001), groups (Balkundi and Harrison 2006; Brass 1981; Mayo and Pastor 2005; Reagans and Zuckerman 2001; Sparrowe et al. 2001) and organizations (Ahuja 2000; Podolny and Barron 1997; Powell et al. 1996; Raz and Gloor 2007; Uzzi 1996).

Based on this stream of research, our hypotheses are structured in two parts. In the first part, we propose two hypotheses that examine the effects of structure and position of a university alumni network on the success of their entrepreneurial activities. In the second part, we present two hypotheses regarding the structure and position of individual entrepreneurs as an antecedent for their success.

2.1. Performance of the alumni network

On the university level, we analyze the cohesiveness of the social network of alumni of a university. Motivated by research by Mayer & Puller (2008), who by analyzing friendships networks on Facebook of university students found that same university, race, and interests were the strongest predictors of friendships, we expect to find cliques of alumni from the same university in the German founder network. We would therefore expect similar behavior for groups of entrepreneurs made up of old-boys networks.

Actors in decentralized networks are typically more interdependent, which leads to an increased willingness to cooperate. With respect to the effect of group density on performance, Reagans and Zuckerman (2001) note that tighter group density leads to improved performance. This result is also confirmed by Balkundi and Harrisons" (2006) meta-analysis. One theoretical argument in favor of this is that the propagation of implicit knowledge is more difficult in sparse workgroups (Hansen 1999). Additionally, a large number of interactions between team members is indicative of mutual dependencies (Sparrowe et al. 2001) which in turn promote collaboration and thus improve the group's performance (Molm 1994). Hence we propose,

H1: The higher the cohesiveness of an alumni network defined as the ratio of internal links to external links, the higher the probability for its aggregated entrepreneurial success.

Authors such as Levi et al. (1954) conclude that increasing centralization of group leaders improves the performance of the groups. In their analyses, Raz and Gloor (2007), Cross and Cummings (2004), and Balkundi and Harrison (2006) also conclude that teams that occupy a central position within the inter-group network, or led by a group manager with a central position in the intra-group network, perform better. Another study has shown that network efficiency is measured on the basis of the aggregate centrality of agents (Schweitzer et. al. 2009). The results of these studies might be explained by the fact that more centralization in the group network provides access to relevant resources. Hence, we propose

H2: The higher the centrality of a university alumni network, the higher the probability that the aggregated entrepreneurial performance of the alumni network is comparatively high.

2.2. Performance of entrepreneurs

It has been shown that CEOs of startups are more successful if they communicate more with their peers (Raz & Gloor, 2007). In particular, Raz & Gloor (2007) analyzed 100 software startups in Israel in 1997, before the eBusiness bubble burst. In 2004 they checked back on which startups were still around. They found that the communication intensity of the CEOs with their peers significantly correlated with the probability of survival of the CEO‟s startup. Baum, Calabrese & Silverman (2000) obtain a similar result when analyzing the Canadian Biotech industry, where they found that the chances of success of a startup increased with the size of its alliance network at the time of founding. Cummings & Cross (2003) examined 182 work groups in a global organization and found that certain network structures are related to performance. Uzzi (1996 & 1997) was also studying social structures and the consequences of embeddedness for the economic performance of organizations. In general, he found that up to a certain threshold embeddedness has positive effects on economic performance. Hence we propose,

H3: The higher the centrality of an entrepreneur, the higher the probability that she or he is successful in comparison with other entrepreneurs.

On the individual level, it has already been shown that people connecting structural holes are more successful (Ahuja 2000, Burt 2004). On the other hand, we also speculate that people well embedded into the old-boys network of their university are more successful. Murray (2004) suggests that academics who start biotech firms use their social capital to recruit collaborators through their local laboratory networks. Gulati (1995) found that business relations commonly grew from prior friendship ties. McPherson, Smith-Lovin & Cook (2001) also studied homophily in social networks. They argue that people‟s personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. The concept of homophily applies to offline and online social networks. Based on extensive research on the success of “old-boys networks” (Simon & Warner, 1992) it has been shown that employees recruited through old-boys networks get higher salaries and are more successful on the job.

H4: The better connected an entrepreneur is with peers of her or his alumni network compared to links with outside peers, the higher the probability that she or he is successful.

3. The Empirical Study

To test the proposed hypotheses, we automatically collected the social network of business relationships of students, entrepreneurs, and executives as captured on Xing (http://www.xing.com). Xing is the leading German language business networking Web site, similar to LinkedIn. People on Xing have the option of either hiding or disclosing their profile to the outside world, as well as of hiding or disclosing their friends. If people choose to make their profile publicly accessible, while also showing their friends, this information is accessible to search engine crawlers. According to its own Web site (July 2009) Xing has over 7 million active user profiles.

For our analysis we focus on 12 German universities which can be classified into three groups: (1) University of Cologne, HU Berlin, University of Hamburg, University of Hannover, and University of Mannheim are among the largest universities in Germany. (2) In addition, we included two well-respected privately organized business schools, European Business School of Oestrich-Winkel (EBS) and WHU Otto Beisheim School of Management. (3) Finally, we added five of the newly selected elite institutions of Germany, LMU Munich, FU Berlin, RWTH Aachen, TU Munich, and University of Karlsruhe.

3.1. Data Collection and Sample

For our research we systematically parsed the publicly accessible alumni profiles of the above universities and were thereby searching for keywords such as Chief, Inhaber, Besitzer (owner), Unternehmer (entrepreneur), Jungunternehmer (junior entrepreneur), Gesellschafter (shareholder), Geschäftsführer (CEO), Geschäftsführender (CEO), Gründer (founder), Teilhaber (Co-owner), Enterpriser, Entrepreneur, and Startup for users from the 12 selected German universities. Overall we collected 654,193 users and 4,456,393 relations from Xing as of April 2009. Out of this large data sample we retrieved 15,143 founders and entrepreneurs with 232,390 relations whose profile matched the keywords (see table 1 for detailed data).

University	*Total students Ø 2004-07*	*Graduates Ø 2004-07*	*Graduating quotient*	*Alumni (Xing sample)*	*Founders (Xing sample)*	*Founder quotient*
U Cologne	45158	5019	11%	7826	2210	28%
LMU Munich	43722	6025	14%	6504	2726	42%
U Hamburg	37518	4982	13%	9128	2526	28%
FU Berlin	33646	4356	13%	6172	1608	26%
HU Berlin	29570	3683	12%	1650	383	23%
RWTH Aachen	29441	2960	10%	5769	1266	22%
U Hannover	22144	2650	12%	3500	857	24%
TU Munich	21237	3740	18%	3076	1262	41%
U Karlsruhe	17579	2089	12%	3577	821	23%
U Mannheim	11089	1380	12%	3562	826	23%
EBS	1270	285	22%	554	173	31%
WHU	444	158	36%	658	196	30%

Table 1. Basic data for all 12 universities

In addition to the social network data, we gathered data on the number of inscribed students and the number of students graduating each year from 2004 to 2007 of the 12 universities (left columns of table 1). The 3 columns on the right of table 1 list the basic data we collected from Xing.

3.2. Measures

This section describes the measures we used for our analysis. The paragraphs 3.2.1-3.2.4 present our estimate of the performance of both university alumni networks and entrepreneurs. In this context we define the following new measures: Graduating quotient, founder quotient, economic impact of founder network, and economic impact per founder. The paragraphs 3.2.5 and 3.2.6 define how we computed the network structure of the alumni networks and the network position of each actor.

3.2.1. Graduation Quotient

As a first performance measure for a university we take the "graduating quotient", i.e. the number of students graduating per year among all students registered. According to this measure the two private schools WHU and EBS are the leaders. Since private universities usually offer shorter durations of study e.g. due to less students per class and a tighter organization of the study schedule this comes as no surprise. However, the two large state universities in Munich (LMU Munich and TU Munich) are also very efficient in guiding their students to graduation in a short time.

3.2.2. Founder quotient

Our second performance measure is called "founder quotient". It is calculated as the percentage of company founders and entrepreneurs among all alumni of a university (based on the Xing data). The values for each University can be found in table 1. Figure 1 illustrates that there is also a (non-significant) correlation between the efficiency of a university, defined as the graduating quotient, i.e. the percentage of students graduating per year, and the propensity of alumni of a university to found businesses, defined as the founder quotient, i.e. the percentage of alumni in Xing listing themselves as founders ($R=0.30$, $p=0.34$).

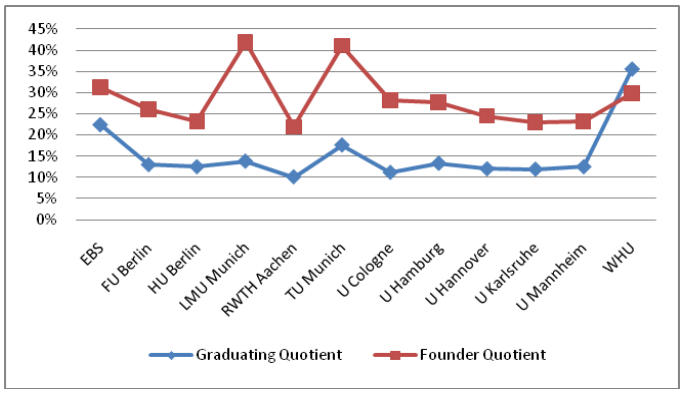

Figure 1. Correlation between graduating and founder quotient

3.2.3. Economic Impact

To measure individual and university success, we randomly picked 80 founders from each university. We then looked at the characteristics of the companies they started. We put the companies into five categories, based on number of employees (1 employee, 2-10, 10-50, 50-200, 200-1000) (Frank-Bosch 2003) and calculated an average annual revenue based on annual average income for these categories (Mercer 2009) (€33k/employee, €33.88k/employee, €34.76k/employee, €35.64k/employee, €36.52k/employee). It has been found elsewhere that the larger the size of the company, the higher the average income of the employees (Frank-Bosch 2003). In addition we looked at the legal form of the startup, adding the amount of equity required to register the company (€50k for an incorporated company (AG), €25k for a limited partnership (GmbH)).

University	*Relative Economic Impact*	*Economic Impact per Founder*	*Total Economic Impact of Founder Network*
LMU Munich	47,994,000 €	599,925 €	1.4 Billion €
TU Munich	43,867,280 €	548,341 €	622 Mio €
WHU	42,789,560 €	611,279 €	108 Mio €
U Hamburg	37,886,760 €	473,585 €	1 Billion €
U Cologne	36,278,360 €	518,262 €	1 Billion €
EBS	32,152,040 €	434,487 €	68 Mio €
FU Berlin	27,643,400 €	337,115 €	487 Mio €
U Karlsruhe	24,815,880 €	310,199 €	229 Mio €
U Mannheim	20,450,960 €	296,391 €	220 Mio €
HU Berlin	15,252,040 €	186,000 €	64 Mio €
RWTH Aachen	13,398,520 €	191,407 €	218 Mio €
U Hannover	12,692,480 €	158,656 €	122 Mio €

Table 2. Economic impact of university on region per year

Table 2 lists the annual contribution of each university to the German economy based on the calculations described above. The column “Relative Economic Impact” shows the average economic contribution of the 80 founders of each university computed according to the above formula. The column “Economic Impact per Founder” is the share of an individual founder. It is calculated by dividing the relative economic impact of a university through the amount of distinct companies founded by the 80 founders. For relatively small universities like WHU and EBS in our 80 people sample there is more than one founder involved in the same company which means that there are less than 80 distinct companies. This is not the case for large universities like e.g. Humboldt University Berlin, where the 80 founders founded 80 distinct companies. Nevertheless, even when this is taken into account, the economic impact per founder is still much higher at WHU and LMU Munich, because their startups are much more successful.

The “Total Economic Impact of Founder Network” in table 2 contains an estimate of what the founders of each university that we identified on Xing contribute to the GDP of a region, computed by multiplying the number of founders identified out of the Xing sample from table 1 with the economic impact per founder from table 2. Obviously, the larger a university, the higher the total number of founders and entrepreneurs and thus the higher the total economic impact. E.g. University of Cologne has a total economic contribution of 1 Billion €, although it

is only ranked on 4th place when looking at revenue generation per founder. Nevertheless, LMU Munich stands out, because although being a large state university, it is also second best in revenue generation per graduate, leading to a staggering contribution of 1,4 Billion € per year. However, because these are absolute numbers and are not directly comparable we only use the relative economic impact and the economic impact per founder for further analysis.

3.2.4. Company Success

To better understand the interrelationship between individual success and social networking behavior, we also looked at the accomplishments of the 80 entrepreneurs whose companies we analyzed.

Level of success	Description
1	Company bankrupt / web site not existing / side business < 1 year
2	Company in business < 5 years / side business
3	Small or medium size business > 5 years / main income / successful
4	Medium size / family business/ stable / very successful
5	Large company / highly successful projects / external funding / rewards

Table 3. Success categories for individual entrepreneurs

Table 3 lists the criteria we applied to rank entrepreneurial success of the 80 individuals we had picked at random from our dataset on a scale from 1 to 5. We read the Web sites and checked business accomplishments of each startup in business databases and assigned each individual to a success category. Table 4 shows the summary of the number of entrepreneurs from the 12 universities in each of the five success categories.

University	Level of Success				
	1	2	3	4	5
LMU Munich	0	7	33	33	7
TU Munich	0	19	48	10	3
WHU	3	6	32	26	3
U Hamburg	0	7	40	29	4
U Cologne	1	14	35	18	2
EBS	0	3	40	27	4
FU Berlin	0	13	39	28	2
U Karlsruhe	0	10	37	29	4
U Mannheim	3	17	36	12	1
HU Berlin	0	22	44	14	2
RWTH Aachen	2	16	37	15	0
U Hannover	0	14	51	14	1

Table 4. Number of entrepreneurs from the 12 universities in each of the five success categories

3.2.5. Network structure and group measures

We computed the (normalized) group degree centrality (GDC), (normalized) group betweenness centrality (GBC) and the ratio of nodes to edges for each of the 12 alumni networks and for the overall network consisting of all alumni networks (Everett & Borgatti 2005, Wasserman & Faust 1994).

For each university we calculated in-group (tribal network), and out-group (full network) statistics. We retrieved the number of actors and edges for the full networks by considering all links from alumni of a university to people from other universities or external institutions. We also calculated the ratio of nodes to edges for all tribal and full university alumni networks. Note that the lower this value, the higher the degree of connectivity of the considered network because there are proportionally more edges connecting the actors. This gives us a simplified measure of

how strongly connected the actors in the different university networks are. Cummings & Cross (2003) use a similar measure. They study the implications of different network structures on group performance and argue that more integrative structures will be related to higher performance.

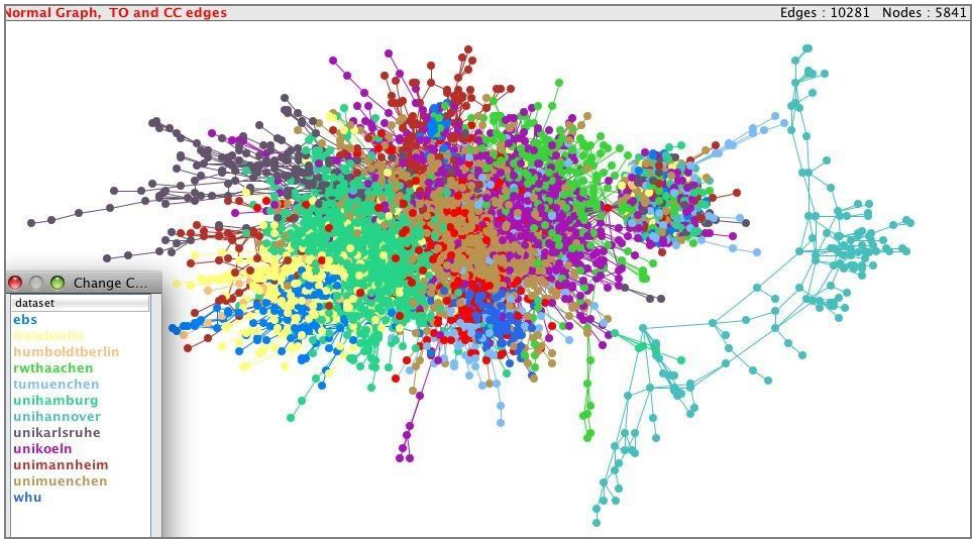

Figure 2. Full network of all founders with more than 5 links (n=5841); light brown dots are alumni of LMU Munich

Figure 2 displays the full founder network of all 12 universities as well as their external friends where each actor has at least 5 connections. Note that out of all actors in our analysis, only 15,143 are founders and alumni from one of the 12 universities, while 130,390 are their Xing friends, either alumni from the 12 universities or from other external institutions. On the university level there is a major cluster of alumni of LMU Munich in the center of the network.

3.2.6. Network structure and individual network position

Figure 3 displays the full network of the EBS alumni, with the alumni shown as dark blue dots, all others as light blue dots. While the dark blue actors are much fewer in number, they are the most connected, and have the highest betweenness.

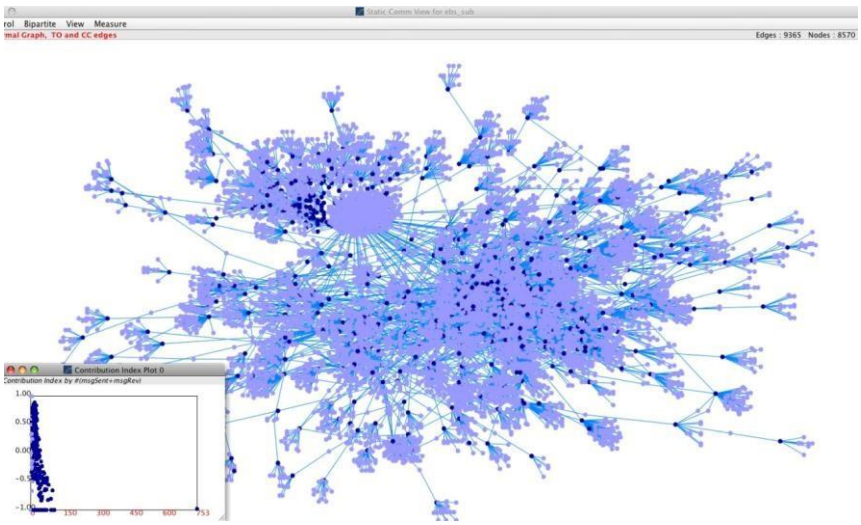

Figure 3. Social network of EBS alumni and their friends (dark blue=EBS alumni, light blue=non-EBS actors), lower left shows contribution index

Motivated by earlier research, where we had compared the full and the group-internal network (Joo et. al 2005) to find the most influential members of a group, we also analyzed the internal – or as we call it in this paper – the "tribal network" of university alumni only (figure 4).

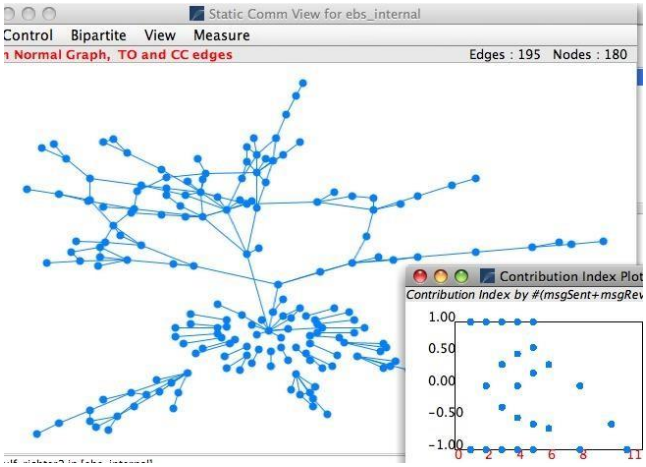

Figure 4. Tribal Social Network of EBS alumni

4. Results

In this section we present the results from our hypotheses testing. We first look at the results on the university level, discussing findings of interest to university presidents to increase entrepreneurial capabilities of their student bodies as well as to students with entrepreneurial interests, to choose the university best suited to their needs.

4.1. Network structure and group performance

We analyze the impact of the founder network on economic performance on two dimensions: (1) we compare social network metrics of each university network (such as e.g. shown in figure 3 and 4) with university-wide performance metrics, and (2) we measure how "tribal" or "incestuous" the old-boys-network of each university is by comparing the "tribal" network (in figure 4) against the "full" network pictured in figure 3.

(1) First we are looking at the social network metrics of the networks. As table 5 shows, there is significant correlation between the tribal network metrics and some of our four measures of performance (Graduating Quotient, Founder Quotient, Relative Economic Impact, and Economic Impact per founder). Interestingly, there are no significant correlations between social network metrics of full university networks and measures of performance.

The higher tribal group betweenness centrality and tribal group degree centrality, the higher the graduating quotient, i.e. the faster students are in getting their degrees (R=081***, R=0.93***, respectively). This means that a centralized university alumni network, which has a few superconnectors, is an indicator for a university that gets out students fast. We speculate that the type of person who has "superconnector" characteristics, i.e. a person with many Xing-friends, is more attracted to a private university with high graduating quotient such as WHU and EBS. There is also significant correlation between tribal group betweenness centrality and economic impact per founder (R=0.52*). The question now is: do superconnectors breed success, or does success breed superconnectors? To put it in other words: is an alumni network, which has superconnectors, better in creating startups that are successful? It could also be that successful entrepreneurs will just get many friends, as everybody will want to be associated with them? We are not yet in a position to give an answer to this question, but let us wait until we have looked at characteristics of individual entrepreneurs in section 4.2.

(2) As a second step of analysis of group performance we measure the openness of the old-boys network to the outside world. We determine the strength of the tribe of alumni of a university – the degree of "tribeness". We define "tribeness" as the ratio of the number of actors and edges within the old-boys network to the number of actors and edges in the outside (external) network of a university:

Node Tribeness = # tribal nodes/# external nodes *Edge Tribeness = # tribal edges/# external edges*

As a metric for the density of the tribal network we also measure the ratio of tribal nodes to tribal edges. The smaller this ratio, the higher is the connectedness of the tribal network. As table 5 illustrates, we get significant negative correlation for these metrics. This means that the more densely connected the actors in the tribal network

are, the better is the university in getting students out, and in creating startups that are financially successful. This would again imply that the university should invest into building a cohesive tribe!

In table 5 we list the tribeness values based on nodes and edges for all universities. Note that the higher the node tribeness, the less external actors are connected to members of the tribal network and thus the more the network has the characteristic of a tribe. The edge tribeness is different in the sense that the amount of "edge tribeness" is determined by how many links go from members within the tribal network to the outside world.

We observe that both measures of tribeness have strong positive significant correlation with the relative economic impact of the university and the impact per founder. These findings indicate that it pays off to be an "incestuous" community and that creating a university with a strong in-group feeling promoting strong bonding among alumni is a means to success. Edge tribeness also positively correlates with the graduating quotient, meaning that having more links within the tribe than with people outside the university promotes fast graduation.

University	**"tribal"** (in-group) **BC**	**"tribal"** (in-group) **Degree**	**Ratio Tribal Nodes/ Tribal Edges**	**Node Tribeness**	**Edge Tribeness**	**Group Betweenness of full network**	**Group Degree of full network**	**Ratio Full Nodes/ Full Edges**
EBS	0.0999	0.0499	0.9231	0.0210	0.0213	0.2591	0.0876	0.9151
FU Berlin	0.0424	0.0286	1.2441	0.0181	0.0133	0.1072	0.0169	0.8984
HU Berlin	0.0059	0.0367	1.7333	0.0053	0.0031	0.1599	0.0448	1.0199
LMU Munich	0.0946	0.0146	1.0397	0.0273	0.0224	0.1781	0.0295	0.8350
RWTH Aachen	0.1557	0.0304	1.0871	0.0219	0.0193	0.0921	0.0182	0.9410
TU Munich	0.0564	0.0431	0.9472	0.0248	0.0248	0.0768	0.0091	0.9210
U Cologne	0.0792	0.0241	1.0455	0.0226	0.0193	0.3124	0.0998	0.8787
U Hamburg	0.0917	0.0137	0.9834	0.0312	0.0264	0.0656	0.0166	0.8101
U Hannover	0.0075	0.0237	1.3830	0.0212	0.0149	0.0806	0.0157	0.9602
U Karlsruhe	0.0526	0.0253	1.2147	0.0189	0.0150	0.1206	0.0316	0.9473
U Mannheim	0.0328	0.0362	1.2345	0.0134	0.0105	0.2519	0.0613	0.9585
WHU	0.3510	0.1848	0.5246	0.0258	0.0421	0.1684	0.0533	0.8230
Correlations								
Graduating Quotient	0.81***	0.93***	-0.72***	0.271	0.77**	0.13	0.24	-0.43+++
Founder Quotient	0.14	0.06	-0.46+	0.53*	0.46+	0.02	-0.08	-0.49*
Relative Economic Impact	0.42++	0.28	-0.71***	0.67**	0.70**	0.15	0.11	-0.80***
Economic Impact per founder	0.52*	0.38	-0.78***	0.66**	0.75***	0.26	0.23	-0.81***

Table 5. Graduating and founder quotient and social network statistics for both the tribal and the full alumni networks for the 12 universities (*p≤0.1; **p<0.05; ***p<0.01; + p=0.13 ++p=0.17,+++p=0.16)

To resume, we have proven hypothesis H1 – the more tribal an alumni network is, the higher the economic output of the university"s founders, and hypothesis H2 – the more centralized a university network is, the higher the success of the university"s entrepreneurs.

4.2. Network structure and individual performance

To compare the network position of entrepreneurs with the economic performance of their company, we are analyzing their individual network structure properties. In addition to actor degree centrality, we defined three additional metrics based on degree centrality, namely (1) Individual Tribe Factor, (2) Individual Weighted Tribe Factor, and (3) Individual BC Tribe Factor.

The *Individual Degree Centrality* for an actor is computed as the number of links that the actor shares with actors of the overall network (Wasserman & Faust, 1994).

The Individual Tribe Factor is determined as the ratio of the number of links a user has with actors of her or his alumni network (in-group) to the number of links the user has with people outside of her or his alumni network (outgroup):

Individual Tribe Factor = In-group Degree (= tribal links) / Out-group Degree (= external links)

This measure allows us to assess how much an entrepreneur is tied to her or his alumni network compared to her or his connections with entrepreneurs who did not study at her or his university.

The combination of these two measures allows us to calculate the Individual Weighted Tribe Factor, to also include the overall size of an individual‟s network:

*Individual Weighted Tribe Factor = individual degree centrality * (# tribal links/# external links)*

We also define a tribe factor based on the relationship of the betweenness centrality (BC) of an individual in the tribal network and the BC of an individual in the external network (Individual BC Tribe Factor). The BC is a precise measure of a founder‟s position in the network, therefore we expect to get a good predictor of an individual‟s membership to a tribe. This means, the more an individual invests into the old-boys network compared to the outside network, the larger is her or his tribal BC and the higher her or his affiliation with the own tribe.

Individual BC Tribe Factor = tribal betweenness centrality / external betweenness centrality

Assessing individual success, we found that individual degree, weighted tribe factor, and betweenness centrality tribe factor and success are positively correlated (table 6). This means that a successful entrepreneur has proportionally more links with other alumni from her/his alma mater than with outside people.

The highly significant correlation between individual degree, i.e. total links of an entrepreneur and her or his success in running the business means that having many Xing friends is an indicator of business success. But as already remarked above, this correlation alone does not answer the question of causality. Based on previous research however (Raz & Gloor, 2007), where we found that startups that have larger informal communication networks increased their chance to survive external shock, we speculate that having many friends in the online world is indeed supportive of later business success.

Level of success	*N*	*Individual Tribe Factor*	*Individual Degree*	*Individual Weighted Tribe Factor*	*Individual BC Tribe Factor*
1	8	0.008578431	4.625	0.039675245	0
2	150	0.042017913	17.95903212	0.754601042	0.086672447
3	472	0.049580305	19.71417444	0.97743479	0.136807269
4	254	0.045653527	27.49223998	1.255117729	0.21345296
5	34	0.028472560	35.09325397	0.999194773	0.19420585
Correlation	918	0.41	0.98***	0.83*	0.94**

Table 6. Average tribe factors for all manually examined actors of all universities (N=918) (*p≤0.1, ** p≤0.05, *** p≤0.01)

The individual BC tribe factor correlates on a highly significant level with our five success levels, which means that the higher a founder‟s embeddedness with the own tribe, the more successful she or he is in building up the business.

It could be, however, that there is an optimum after which investing too much into the tribal network becomes counter effective. We speculate that tribeness has some common characteristics with the concept of embeddedness as studied by Uzzi (1996 & 1997). Uzzi argues that the positive effect of embeddedness that firms organized in tightly connected networks have higher survival chances reaches a threshold, after which the effect reverts itself. Extrapolating Uzzi‟s results would imply that there is a threshold after which being a tribe does not pay off anymore. In analogy to his findings we observe that the most successful entrepreneurs in table 5 (on success level 5) have somewhat lower values for all individual tribe factors. This in other words means, that the most successful founders have proportionally somewhat more links to the outside world than within their own tribes than founders on success level 4.

To resume, we have proven hypotheses H3 – the more online friends an entrepreneur has, the more successful she or he is, and H4 – the more tribal an entrepreneur is, the more successful she or he is.

5. Discussion

The popularity of online social networking is unbroken. People use these sites to connect with family, friends, and business contacts. For many people, particularly in the generation of the 15 to 30 year olds, it is a substitute for email or phone. Based on previous work in the same research field (e.g. Raz & Gloor 2007, Uzzi 1996 & 1997, Cummings & Cross 2003) we were studying networking structures of entrepreneurs and founders in cyberspace to predict an entrepreneur"s success. Through analysis of the largest German business social networking platform Xing we could identify clusters of entrepreneurs at 12 major German universities. Our main goal was to find out how their online networking behavior affects their success of founding new businesses. We divided our analysis into two parts: In the first part, we verified two hypotheses that examine the effects of the structure and position of a university alumni network on the success of their entrepreneurial activities. In the second part, we presented two hypotheses regarding the structure and position of individual entrepreneurs as an antecedent for their success.

"Birds of a feather flock together". Many studies dealt with this phenomenon and found out that social groups are not random samples of people (e.g. Mayer & Puller 2008, McPherson, Smith-Lovin & Cook 2001, McPherson, Popielarz, Drobnic 1992). It has also been shown that the intensity of communication of these groups has an impact on performance (Raz & Gloor 2007). Most of the work was done with offline social networks, e.g. by studying work groups or organizations. In this study we could identify similar effects in an online social network (Mayer & Puller 2008).

We assumed that there are certain structural properties of these networks that will explain success. Group betweenness centrality and connectivity of the alumni network were strongly correlated with the efficiency of a university. We measured the efficiency of a university by looking at the average number of students graduating each year depending on the average number of total students inscribed. We found that universities with hierarchically organized alumni networks and higher degree of internal connectivity were faster in getting their students out.

We also found that university alumni networks that were successful in founding startups – measured by their average economic contribution – are organized as tribes. We found that their tribeness, the strength of their internal cohesiveness or their negative degree of openness to external actors correlates strongly with their economic success. For a university this could mean that it should foster and encourage students to build up more and closer connections with alumni. Porter et. al. (2005) found that nearly all university-educated founders retain some form of affiliation with their universities after successfully starting their business. But as we have found there might be a threshold of embededdness (Uzzi 1996 & 1997) after which the positive effect of connecting to the own people might tamper off. For growing an environment for the most successful business leaders, the founder also needs connections to external people and institutions to a certain extent. In our own data we especially found this while analyzing success on an individual level. For the most successful founders (level 5 in our analysis), their tribe affiliation was slightly lower than for founders on the level right below. Proportionally, they were having somewhat more links to external actors than to people within their university.

Findings on university level and on individual founder level show that the more a founder is embedded in her/his own tribe, the more successful will the business be. Universities whose alumni prefer friends from the same university seem to be more successful in creating new businesses and generating higher economic contribution per startup founder.

6. Limitations

The main question is if our sample is relevant of the entire population of German founders. On the one hand one can argue that there is a significant proportion of particularly older, forty to sixty year old founders who do not have a profile on Xing, LinkedIn, or Facebook. However, the online world has become a mirror of the real world. Trendsetters such as founders of new businesses use online media to communicate and stay in touch. These entrepreneurs, whether they are in the Web savvy age group of twenty to forty year olds or older, have a high likelihood of using tools like Xing to stay in touch.

One can also make the argument that our technique of sampling the entrepreneurs by choosing 80 profiles randomly per university distorts our results because the likelihood of finding multiple founders of the same business is higher for small business schools with smaller overall number of students.

7. Future Work and Conclusion

It will be necessary to complement our findings with more studies of the offline world. For example, a complementary offline survey of university alumni would give us a more comprehensive overview of the embeddedness of an alumnus in her/his social network.

We also intend to further analyze existing activities of universities of educating their students in entrepreneurship or starting a business and put this in relationship to our metrics of economic success.

Looking at the content or type of information flow between people in an online social networking platform could reveal insights about the kind of relationship and whether there are differences among ties (e.g. casual acquaintance against close collaboration), it could be that not all types of ties support the same level of success (Aral & Van Alstyne 2007). Usually ties in an online social network and especially on Xing do not hold such information directly. However, it might be possible to extract information from the profiles of the connected actors and derive the type of the relationship through a content analysis of the affected profiles.

Nevertheless, we have shown that it pays to have many business contacts also in the online world, and to choose these contacts well – the better you know them as alumni of your own university, the more successful your business will be.

References

Aguillo, I. F.; Granadino, B.; Ortega, J.L.; Prieto, J.A. (2006). "Scientific research activity and communication measured with cybermetrics indicators". Journal of the American Society for Information Science. Volume 57 Issue 10, Pages 1296 – 1302. May 2006

Ahuja, G. (2000). "Collaboration networks, structural holes, and innovation: A longitudinal study". Administrative Science Quarterly 45(3) 425–455.

Allen, T.; Raz, O.; Gloor, P. (2009). "Does Geographic Clustering Still Benefit High Tech New Ventures? The Case of the Cambridge/Boston Biotech Cluster". MIT ESD-WP-2009-01 working paper 2009

Allen, T.J. (1984). "Managing the Flow of Technology", MIT Press, Cambridge, MA.

Aral, S.; Van Alstyne, M. (2007). "Network Structure & Information Advantage", in Proceedings of the Academy of Management Conference, Philadelphia,

PA. Balkundi, P. and D.A. Harrison. (2006). "Ties, Leaders, and Time in Teams: Strong Inference About Network Structure's Effects on Team Viability and

Performance." Academy of Management Journal 49(1): 49-68.

Baum, J.; Calabrese, T.; Silverman, B.S. (2000). "Don"t go it alone: Alliance network composition and startups" performance in Canadian biotechnology". Strategic Management Journal. 21 267–294.

boyd, d. (2008). "Why Youth. Social Network Sites: The Role of Networked Publics in Teenage Social Life". Youth, Identity, and Digital Media. Edited by David Buckingham. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, 2008. 119–142.

boyd, d.; Ellison, N. (2008). "Social Network Sites: Definition, History, and Scholarship". Journal of Computer-Mediated Communication 13 (2008) 210–230.

Brown, J.S.; Duguid, P. (1991). "Organizational Learning and communities of practice: toward a unified view of working, learning and innovation", Organization Science, Institute for Operations Research and the Management Sciences, 2:1, 40-57.

Burt, R. (2004). "Structural Holes & Good Ideas". American Journal of Sociology, (110): 349-99.

Brass, D.J. (1985). "Men"s and women"s networks: a study of interaction patterns and influence in an organization". Academy of Management Journal 28:327-43.

Castells, M. (2000). "The Rising of the Network Society", Blackwell Publishers Ltd, Oxford.

Cothrel, J.; Williams, R.L. (1999). "On-line communities: helping them form and grow", Journal of Knowledge Management, 3(January), 54-60.

Cummings, J.; Cross, R. (2003). "Structural properties of work groups and their consequences for performance." Social Networks, 25(3): 197-210.

DiMaggio, M.; Gloor, P.; Passiante, G. (2009). "Collaborative Innovation Networks, Virtual Communities, and Geographical Clustering". International Journal of Innovation and Regional Development, Vol 1, No. 4, 2009, pp. 387 – 404

Ellison; Steinfield; Lampe (2007). "The Benefits of Facebook „„Friends:"" Social Capital and College Students" Use of Online Social Network Sites". Journal of Computer-Mediated Communication 12 (2007) 1143–1168.

Everett, M.G.; Borgatti, S.P. (1999). "The centrality of groups and classes", Journal of Mathematical Sociology 23 (3) (1999), pp. 181–201.

Everett, M.G.; Borgatti, S.P. (2005). "Extending Centrality" in Carrington, et. al. (2005) Models and Methods in Social Network Analysis.

Forschungsgruppe Wahlen (2009) http://www.fgw-online.de/Umfragen_und_Publikationen/Internet-Strukturdaten/web_II_09.pdf

Frank-Bosch, B. (2003): "Verdienststrukturen in Deutschland: Methode und Ergebnisse der Gehalts- und Lohnstrukturerhebung 2001". Wirtschaft und Statistik 12 (2003), 1137-1151.

Gloor, P.; Krauss, J.; Nann, S.; Fischbach, K.; Schoder, D. (2009). "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis". IEEE Conference on Social Computing (SocialCom-09), Aug 29-31, Vancouver, 2009.

Gloor, P.; Zhao, Y. (2006). "Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis". Proceedings of 10th IEEE International Conference on Information Visualisation IV06 (London, UK, 5-7 July 2006)

Gulati, R. (1995). "Does familiarity breed trust? The implications of repeated ties for contractural choice in alliances", Academy of Management Journal 38(1): 85– 112.

Joo, S.; Gloor, P.; Schnorf, S. (2005). "Detection of Power User Patterns Among High School Students in a Mobile Communication Network". Power Users of ICT International Symposium, Costa Rica, Aug.8-10, 2005 Luhmann, N. 1979. Trust and power. Chichester, England: Wiley.

Mayer, A.; Puller, A. (2008). "The old boy (and girl) network: Social network formation on university campuses". Journal of Public Economics, Volume 92, Issues 12, February 2008, Pages 329-347

McPherson, J.M.; Smith-Lovin, L.; Cook, J. (2001). "Birds of a feather: Homophily in Social Networks."Annual Review of Sociology 27: 412-444.

McPherson, J.M.; Popielarz, P.; Drobnic, S. (1992). "Social networks and organizational dynamics". Administrative Sociological Review 57: 153-70.

Murray, F. (2004), "The role of academic inventors in entrepreneurial firms: Sharing the laboratory life." Research Policy 33(4): 643–659.

O"Murchu, I.; Breslin, J.G.; Decker, S. (2004). "Online Social and Business Networking Communities". DERI – Digital Enterprise Research Institute DERI Technical Report 2004-08-11, August 2004

Jones, S.; Fox, S. (Jan 28, 2009). Pew Internet Survey (2009) Generations Online in 2009 (http://www.pewinternet.org/Reports/2009/Generations-Online-in2009.aspx, retrieved July 20, 2009)

Porter, K.A.; Bunker Whittington, K.C.; Powell, W.W. (2005). "The institutional embeddedness of high-tech regions: Relational foundations of the Boston biotechnology community". S. Breschi & F. Malerba (Eds.), Clusters, Networks, and Innovation: 261-296. Oxford, UK: Oxford University Press.

Raz, O. Gloor, P. (2007). "Size Really Matters - New Insights for Startup's Survival". Management Science, February 2007

Romanelli, E. (1989). "Environments and strategies of organization start-up: Effects on early survival". Administrative Science Quarterly 34(3) 369–387.

Saxenian, A. (1994). "Regional Advantage: Culture and Competition in Silicon Valley and Route 128". Cambridge, Harvard University Press, MA.

Schilling, M.A. & C.C. Phelps (2005). “Interfirm collaboration networks: the impact of small world connectivity on firm innovation”. Management Science, 53 (7), pp. 1113-1126.

Schweitzer, F.; Fagiolo. G.; Sornette, D.; Vega-Redondo, F.; Vespignani, A.; White, D.R. (2009). “Economic Networks: The New Challenges”. Science. Vol 325. 24.July 2009, pp. 422-425

Simon, C.J.; Warner, J.T. (1992). “Matchmaker, Matchmaker: The Effect of Old Boy Networks on Job Match Quality, Earnings, and Tenure”. Journal of Labor Economics, Vol. 10, No. 3 (Jul., 1992), pp. 306-330

Uzzi, B. (1996). “The sources and consequences of embeddedness for the economic performance of organizations: The network effect”. American Sociological Review 61(4) 674–698.

Uzzi, B. (1997). “Social structure and competition in interfirm networks: The paradoxon of embeddedness”. Administrative Science Quarterly, 42: 35-67.

Wasserman, S.; Faust, K. (1994) “Social Network Analysis”. Cambridge University Press.

White, D. R.; Houseman M. (2002). “The navigability of strong ties: Small worlds, tie strength and network topology”. Complexity 8(1) 72–81.

Contribution IV

Nann, S.; Krauss, J. and Schoder, D. (2013) *"Predictive Power on Public Data – the Case of Stock Markets"* was published in the proceedings of the European Conference on Information Systems (ECIS) as a completed research paper.

7-1-2013

Predictive Analytics On Public Data - The Case Of Stock Markets

Stefan Nann
University of Cologne, Cologne, Germany, stefan.nann@gmail.com

Jonas Krauss
University of Cologne, Cologne, Germany, krauss@wim.uni-koeln.de

Detlef Schoder
University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

PREDICTIVE ANALYTICS ON PUBLIC DATA – THE CASE OF STOCK MARKETS

Nann, Stefan, University of Cologne, Pohligstr. 1, 50969 Köln, Germany, nann@wim.uni-koeln.de

Krauss, Jonas, University of Cologne, Pohligstr. 1, 50969 Köln, Germany, krauss@wim.uni-koeln.de

Schoder, Detlef, University of Cologne, Pohligstr. 1, 50969 Köln, Germany, schoder@wim.uni-koeln.de

Abstract

This work examines the predictive power of public data by aggregating information from multiple online sources. Our sources include microblogging sites like Twitter, online message boards like Yahoo! Finance, and traditional news articles. The subject of prediction are daily stock price movements from Standard & Poor's 500 index (S&P 500) during a period from June 2011 to November 2011. To forecast price movements we filter messages by stocks, apply state-of-the-art sentiment analysis to message texts, and aggregate message sentiments to generate trading signals for daily buy and sell decisions. We evaluate prediction quality through a simple trading model considering real-world limitations like transaction costs or broker commission fees. Considering 833 virtual trades, our model outperformed the S&P 500 and achieved a positive return on investment of up to ~0.49% per trade or ~0.24% when adjusted by market, depending on supposed trading costs.

Keywords: Predictive Analytics, Data Mining, Sentiment Analysis, Financial Markets, Twitter, Social Media.

1. Introduction

Recent academic discourse on information systems speaks of a paradigm shift toward data-intensive computing and "big data"-based studies (Hey, Tansley and Tolle, 2009). A new level of connectedness among peers creates a huge database by providing new ways for the dissemination and consumption of data and ever-easier means of collecting vast amounts of public data from various on- and offline resources, including posts, tweets, Web documents, and news feeds. Evolving data mining technologies and the increasing processing power of today's computers support the desire to appropriately analyze at least parts of today's growing (public) data deluge in real time, and thus tackle the core question of what meaningful information can be derived through algorithmic analyses and what predictive value can be inferred in automated fashion from public data.

Social media data in particular have been the subject of academic research in the recent past (Schoder et al., 2013). Asur and Huberman (2010) investigated the predictive power of tweets for box office returns, analyzing some 2.89 million tweets. Bollen, Mao and Zeng (2010) collected and classified tweets to forecast daily closing values of the Dow Jones Industrial Average. Among many others (Koch and Schneider, 2002; Forster, 2002; Antweiler and Frank, 2004; Wang, Jank and Shmueli, 2008; Xu et al., 2012), these studies represent the research field of predictive power in publicly available data. In their research essay, Shmueli and Koppius (2011) suggest a framework they call "predictive analytics," which is concerned with the assessment of predictive power in empirical research and statistical inference, and propose six roles for its application. This paper adopts the role of "the assessment of the predictability of empirical phenomena" from their framework.

The work presented herein extends previous attempts to assess the predictive power of social media talk by aggregating data from multiple resources (Twitter, eleven online message boards, and traditional news) and considering an extended period of six-month of data. The subjects of prediction are daily stock price movements from the Standard & Poor's 500 index (S&P 500). Most research concerned with stock market predictions based on online data is mainly theoretical in nature and does not take into account real-world limitations such as broker fees, bid/ask differences, and liquidity. To demonstrate the potential practical application of our findings, we describe a simple trading model based on the predictor, considering commission fees, transaction costs, and stock liquidity.

2. Literature Review

This section provides a brief summary of past literature concerned with the predictive power of online data for financial markets and shows how results improved with the progress of time. While older research found that discussion followed market movements, more recent results clearly detect predictive value in online data for stock price changes.

There are two main research streams which are relevant for the scope of this work. The literature can be categorized in works related to sentiment analysis and works related to predictive power of user generated content (UGC). Sentiment analysis is a broad research field and is applied on many different domains (Berger, Della Pietra and Della Pietra, 1996; Pang, Lee and Vaithyanathan, 2002; Whitelaw, Garg and Argamon, 2005; Abbasi, Chen and Salem, 2008; Boiy and Moens, 2009; Choi, Kim and Myaeng, 2009; Lin and He, 2009; Narayanan, Liu and Choudhary, 2009; Mizumoto, Yanagimoto and Yoshioka, 2012; Fang, Datta and Dutta, 2012). The second major research stream relevant for this study relates to works of predictive power of UGC. Although this topic is much broader and applies to many different domains, the following articles focus on making predictions for developments in financial markets based on UGC.

In many studies both streams are tied together since the value of user generated content can be captured better when it is analyzed with automated methods (Antweiler and Frank (2004), Das and Chen (2007), Bollen, Mao and Zeng (2010), Zhang and Swanson (2010), Sprenger and Welpe (2010)).

These authors apply sentiment analysis for extraction of predictive value from UGC and to study its impact on the stock market. Antweiler and Frank (2004) studied the predictive power of online message boards for the stock market by analyzing 1.5 million messages from Yahoo! Finance (http://finance.yahoo.com) and Raging Bull. Applying sentiment analysis, they found that the number of messages is a predictor for stock turnover. However, their model's performance would not deliver a significant return on investment, as plausible transaction costs would be too high. Das, Martinez-Jerez and Tufano (2005) found that sentiment follows stock price returns.

All more recent research applies sentiment analysis to a changing number of messages from a variety of online resources. While Oh and Sheng (2011) looked at a comparably small subset of messages from Stocktwits, Bollen, Mao and Zeng (2010) collected a large amount of ~9.85 million microblog postings from Twitter (http://www.twitter.com). Schumaker et al. (2012) looked at a small sample (9211 news articles) of traditional news articles; the amount of data for other studies falls between these parameters. However, the latest work focuses on a single source of data (mainly Twitter), leaving out other, well-researched sources such as Yahoo! Finance, Raging Bull, or traditional finance news – which may or may not improve the results.

Traditional news evaluated by natural language processing can carry alpha information as well (Cohen and Frazzini, 2008; Schumaker and Chen, 2009; Dion et al., 2011). Alpha information refers to information that is not yet reflected in stock price levels, thus leading to future stock price movement according to the Efficient Market Hypothesis (EMH) (Fama, 1970).

There is a number of studies over the last decade which found predictive evidence of UGC on stock return (Bagnoli, Beneish and Watts, 1999, Tumarkin and Whitelaw, 2001, Jones, 2006, Gu et al., 2006, Das and Chen, 2007, and Sabherwal, Sarkar and Zhang, 2008).

Bollen, Mao and Zeng (2010), Zhang and Swanson (2010), Sprenger and Welpe (2010), Oh and Sheng (2011), and Xu et al. (2012) are examples of recent studies that have found clear evidence for the predictive power of online communication for stock price movements. Oh and Sheng (2011) examined ~72,000 microblog postings from Stocktwits.com, extending over a three-month period, to predict stock price movements. Applying sentiment analysis, they found microblog messages predict future stock price movement. They also briefly evaluated potential return on investments, finding that simple (not adjusted returns) deliver better results than market-adjusted returns.

Our work and that of other researchers hypothesizes that online talk in social media and microblogs has predictive power over future stock price movements. Microblog posts in particular are characterized by strong focus on their subject because they are succinct, happen in nearly real-time, and have high posting frequencies (Xu et al., 2012; Oh and Sheng, 2011; Bollen, Pepe and Mao, 2010; Java et al., 2007).

Andrew Lo (2004) provides a theoretical foundation for the predictive power of public data over financial markets with the Adaptive Market Hypothesis (AMH), as suggested by Brown (2012). Taking behavioral economics and finance research (De Long et al., 1990; Hirshleifer, 2001; Camerer and Loewenstein, 2004; Tetlock, 2007; Xu and Zhang, 2009; Zhang and Swanson, 2010) into consideration, the AMH describes an evolutionary model of individuals adapting to a changing environment via simple heuristics. It provides an explanation for the existence of alpha information and how learning and competition gradually restore market efficiency (Neely, Weller and Ulrich, 2009). Thus, social media, microblog posts, and news could be considered contributors to the competition and learning process that drives prices (Brown, 2012).

This paper contributes by providing a case study of a virtual trading model based on the predictive power of online communication. Our first goal is to demonstrate that it is indeed possible to trade based on an online message board predictor and achieve a positive return on an investment (adjusted by the market). We gathered communication from Twitter, eleven online message boards, and Yahoo! Finance's news stream, thus extending the scope of data utilized in earlier research.

3. Methodology and Sentiment Analysis

For this paper, we collected 2,971,381 messages concerned with stocks of the S&P 500 index during a six-month period from June 1 to November 30, 2011. Table 1 shows the different online sources which we have used to collect the messages.

The first step is to assign these messages to stocks. For that we filtered the dataset by either looking for messages in sub-forums concerned with particular stocks on online message boards or by using Twitter's "cash tag". The cash tag is a stock's ticker symbol with a preceding dollar sign ($). Not considering spam at this point (more details on how we filter spam in the following paragraph) we can rely on that a tweet which contains a cash tag refers to the stock price or anything related to the financial value of the underlying company (e.g. $MSFT for company Microsoft).

In sources other than Twitter stock specific communication can be accessed in sub-forums which exist for each component of the S&P 500 index. In these sub-forums users exclusively discuss topics related to a particular company. For example Yahoo! Finance provides direct access to a stock's sub-forum by appending the ticker symbol to the following hyper-reference: http://finance.yahoo.com/mb/ (e. g. http://finance.yahoo.com/mb/MSFT for company Microsoft). With a similar approach specific subforums can be accessed for all other sources listed in table 1. With this method, each tweet / forum post can clearly be assigned to a single company which is important to ensure only relevant communication is considered when calculating sentiment.

Thus we get a relatively precise assignment of messages to a specific company / stock which helped us to avoid some common name entity conflicts as mentioned in Yerva & Miklós & Aberer (2010). Although we cannot rely with full certainty, we can assume that people talk about the company or company-related issues when posting in the financial discussion board about Apple and not about the fruit apple. Nevertheless, after collecting all messages we applied a spam filter which cleaned our data set. Our self-developed spam filter searches for example for posts which intend to insult other users without contributing relevant information. Most of these posts can be identified by the usage of scurrile and nasty language. Everything related was removed from the data set.

Following the state-of-the-art research as established in the previous section, in the next step we applied sentiment analysis to microblog messages, forum posts, and traditional news using a Naïve Bayes classifier with an adapted bag-of-words in combination with part-of-speech tagging to find negations and spam filtering based on keywords. The basis of the applied sentiment methodology can be found in Krauss and Nann and Schoder (2012). One key finding in this study indicates that the quality of sentiment recognition depends on how specific the sentiment analysis algorithms are adjusted to the analyzed context. The more context-specific the algorithms are designed the higher the quality of sentiment recognition will be. This is determined e.g. by the choice of bag-of-words and the adjustment of part-of-speech tagging. For example, authors use different language and words to write a positive review about a digital camera and say something positive about their favorite stock. In the current study we adjusted the sentiment analysis very specifically to the stock market domain.

For this reason we initially read a few hundred tweets and posts from the available dataset and manually annotated it with positive or negative sentiment. This sample data was used to train our selfdeveloped sentiment algorithm. In the next step we applied the trained algorithm to a newly and not annotated data sample from the available data set to determine the precision of the sentiment algorithm. Also in a manual process we defined lists of positive (e.g. buy, long, call, etc.) and negative (e.g. sell, short, put, etc.) words which resulted in our bag-of-words. During text analysis the algorithm scans the content for these words. We also manually defined specific words for part-of-speech tagging. The word "don't" for example will be used during part-of-speech-tagging, if a user posts "I don't sell my shares", this will be recognized and labeled with positive sentiment since the key word "don't" will give the negative key word "sell" the opposite meaning.

After all, the algorithm calculates a ratio (decimal number) based on the occurrences of positive and negative labels in a tweet or post. The sum of all ratios for all messages of a specific stock represents the aggregated sentiment value which was used to predict the daily stock price.

Messages of all sources were considered with equal weight in our model. It is obvious that different data sources contain different value contributions. Tweets will probably have a lower half-life than traditional news which usually references a longer time period. It is subject of further research to design a process which evaluates these aspects of every single source separately. Please also refer to section 7 Discussion and Research Suggestions.

Source	URL	Number of Messages
Clearstation	http://www.clearstation.com	12,743
Free Realtime	http://quotes.freerealtime.com	610
Hotstockmarket	http://www.hotstockmarket.com	1,818
Investor's Hub	http://investorshub.advfn.com	20,359
Investor Village	http://investorvillage.com	37,180
The Motley Fool	http://www.fool.com	10,587
Raging Bull	http://ragingbull.quote.com	15,331
Silicon Investor	http://www.siliconinvestor.com	21,442
Stockhouse	http://www.stockhouse.com	37,119
The Lion	http://www.thelion.com	5,766
Twitter	http://www.twitter.com	1,801,345
Yahoo! Finance Boards	http://messages.finance.yahoo.com	802,476
Yahoo! Finance News	http://finance.yahoo.com	204,605

Table 1. Data Sources

As not all stocks from the S&P 500 index receive equal attention in social media, there are substantial differences in the average number of messages written each day for different stocks. For instance, Apple Inc. is one of the most discussed equities on the Web and many more messages are posted for Apple Inc. than for other index components. Thus we require an adjustment of sentiment values based on the average number of messages. For this work we chose the simple moving average (SMA) to achieve comparability for equities of differing attention levels. Sentiment values were used as a stock price movement predictor, with positive values indicating an upward movement and negative values indicating a downward movement.

We began calculating our predictor one month after commencing data collection since we used a 30day simple moving average (SMA30) to calculate sentiment values. Thus, predictions of stock price movements were made each trading day from June 1 to November 30, 2011. For each trading day t, the predictor considered sums of positive and negative messages for each stock on the S&P 500 index and weighted them based on the SMA30. This value was used to predict stock price change on day t + 1, predicting an increase in the case of positive values and a decrease in the case of negative values. Sentiment values can assume any value larger or smaller than zero.

$$Sentiment\ Predictor = \frac{\text{No. of positive messages}}{\text{SMA30 of positive messages}} - \frac{\text{No. of negative messages}}{\text{SMA30 of negative messages}}$$

Through weighting current messages in relation to the SMA30, it is possible to compare stocks that have significantly different attention levels and thus strongly differing message averages. We chose a 30-day average because it takes into account enough days to even out positive and negative peaks in communication without being too static in comparison to longer periods. Longer periods would carry the danger of ignoring short-term anomalies in communication – e.g. in the case of earning releases or bankruptcies – which often lead to a strong increase in message numbers.

Each trading day, we determined the level of sentiment (threshold) for each stock where the historical ratio of correct to total predictions is maximized. Sentiment values count as prediction signals only if

the absolute sentiment value lies above the threshold. For instance, for a threshold of two, a sentiment value of one would not be considered a signal to trade. We systematically determined historic prediction ratios by conducting a sensitivity analysis of sentiment thresholds each trading day for each stock. This was done by looking at all past trading days, comparing the sentiment on day t with the stock price change on day t + 1, summing dates on which a positive/negative sentiment on day t corresponded with a positive/negative stock price change on day t +1, and finally building the ratio for each threshold. Results were statistically significant on a level of 0.05 (right-sided significance test). On average, we obtained 10 stocks per trading day that had statistically significant prediction ratios greater than 0.5. We further found that a look ahead of one day delivered the best results, which means that sentiment values of day t predicted stock price changes of day t + 1 with the highest accuracy compared to predictions for day t + 2 or day t + 3.

$$\frac{sum\ correct\ predictions\ (sentiment\ threshold)}{total\ predictions\ (sentiment\ threshold)} \rightarrow max$$

4. Results: Prediction Ratio of Sentiment Signals

Our results show that publicly available data in microblogs, forums, and news on day t have predictive power for stock price changes on day t + 1. We confirm the findings of Bollen, Mao and Zeng (2010), Sprenger and Welpe (2010), and Oh and Sheng (2011) for single source-based predictions and extend their validity to the case of using multiple sources in aggregation. Table 2 displays the overall prediction performance and the performance for each month.

Predictions	All	Ratio Buy Predictions	Ratio Sell Predictions
Entire period	60.38%	60.69%	60.03%
November 2011	57.54%	60.71%	54.74%
October 2011	63.76%	63.16%	64.81%
September 2011	52.63%	48.98%	55.86%
August 2011	67.34%	67.52%	67.18%
July 2011	60.99%	63.38%	56.79%
June 2011	59.73%	59.01%	60.61%

Table 2. Prediction Ratios

The percentage values display the ratios of correct predictions for all analyzed stocks over the entire period from June 1 to November 30, 2011. For example 60.38% means that ~60 percent of all stocks for which sentiment had significant prediction ratios in the past delivered correct predictions in the period considered. In sum, the algorithm made 1,300 predictions over the entire period (126 trading days). Table 2 lists only predictions for rising stock prices through positive sentiment and predictions of falling stock prices through negative sentiment. Both cases do not differ significantly.

5. Trading Model and Model Parameters

To extend the academic body of literature, and especially to go beyond more recent research as illustrated by, for example, Bollen, Mao and Zeng (2010), Sprenger and Welpe (2010), and Oh and Sheng (2011), we demonstrate the potential practical application of our findings. Here we describe a simple trading model based on the predictor, considering commission fees, transaction costs, and stock liquidity. Most research concerned with stock market predictions based on online data is mainly theoretical in nature and does not take into account real-world limitations when considering a return on investments. Although we are taking some of these factors into account we do not propose a complete trading strategy which could be executed on the stock market as is. We did not execute trades under real market conditions.

Our model is based on assumptions and simplifications that are as practical as possible. However, on real-world trading floors and in real trading environments, there are many factors that can influence a trading model that works perfectly in theory. In our opinion, it is most critical to control for large spreads (differences between the bid and ask prices of stocks), the traded volume of a stock (the more a stock is traded, the higher is its liquidity and the higher the chance to buy or sell the stock for the desired price), and broker commissions, which become particularly relevant if a strategy is based on several trades per day (as is ours).

Table 3 shows our three most relevant and important assumptions to simulate a practical trading strategy. For our trading model, we considered only stocks from the S&P 500 index (as described in the section above). The S&P 500 is one of the most important U.S. stock market indices (and probably also in the world), and stocks on the index are traded mostly in the United States. Therefore, all stocks from this index have comparably high trading volumes, which are important to ensure liquidity. To enforce this criterion, we considered only stocks with a trading volume larger than 3 million shares traded on average per day. For these stocks, it is almost guaranteed that shares can be bought in the morning at the opening of the stock market and be sold at the closing bell. Further, we only considered stocks that cost more than $10 on the day the trade is to be executed. This is important to guarantee a relatively small spread.

Criteria	Description
Tradable Stocks	Only trading stocks from S&P 500 index
Stock Volume	> 3,000,000 traded shares/per day on average
Stock Price	> $10 (on the trading day selected)

Table 3. Assumptions for our trading model

The difference between bid and ask prices (spread or transaction costs) is the highest cost factor for a trading model which is based on multiple trades on one day. Transaction costs are commonly expressed as basis points in a finance context. A basis point (bp) is a unit of measure used in finance to describe the percentage change in the value or rate of a financial instrument. One basis point is equivalent to 0.01% (1/100th of a percent) or 0.0001 in decimal form[7]. The difference between the ask/bid price of a stock, which is the price that has to be paid when buying/selling the stock, and the actual stock price typically lies between 10 bp and 20 bp (0.1% to 0.2%) per transaction. However, these values are based on our assumptions that a stock has a trading volume of more than 3 million shares on average per day and is worth at least $10 on the day it is traded. For example, if a stock trades at $10.02 the broker might charge $10.03, which implies trading costs of 10 bp or 0.1%.

For our trading model, we followed a simple trading rule:

- If the sentiment predictor is positive, the strategy is to buy the stock on market open (open long position) and sell the stock (close long position) on market close (on the same day).
- If sentiment predictor is negative, the strategy is to sell the stock (open short position) on market open and buy the stock (close short position) on market close (on the same day).

We assumed that we could buy the stocks for their opening prices and sell them for their closing prices every day[8]. Considering criteria from Table 3 and its intersection with our algorithm's sentiment signals, we obtained about 7 tradable stocks on average each day (833 trades which meet the criteria on 126 trading days). We obtained a daily overall return on investment (ROI) by summing individual ROIs for these stocks. This was done for the entire period from June 1 to November 30, 2011. To simulate the model in a more realistic way, we also adjusted our ROI with market movement in the considered time

[7] http://www.investopedia.com/ask/answers/05/basispoint.asp#axzz1tcGlLhM7

[8] Buying a stock for the opening price and selling for the closing price is a simplification and is not necessary applicable in practice. There are many reasons why it may not be possible to buy the stock for the opening price and sell it for the closing price (e.g., transaction execution time).

period. For this purpose, we used the change of the SPY certificate, which is one of the most liquid and heavily traded titles on the financial market. The trading rule for adjusting ROIs with the market was as follows:

- If the signal is positive, we buy (long) the stock and sell (short) the market.
- If the signal is negative, we sell (short) the stock and buy (long) the market.

The rule of adjusting results with changes of the market is a form of hedging. Concretely, this means that we limited losses if a positive/negative sentiment predictor turned out to be wrong and the stock price actually rose/fell. In many cases, stocks correlate with general market movement, which means that a stock often rises/falls when the market rises/falls. This is the reason that we traded the opposite of the current sentiment predictor on the market. This means that if a stock fell after a positive sentiment, we limited our losses by being short on a falling market. Table 4 (adjusted by market) and Table 5 (unadjusted) summarize the performance of our model with various levels of transaction costs.

6. Results from Trading Model Considerations

Transaction Costs	0 bp	5 bp	15 bp	20 bp	25 bp
Total ROI	196.84%	155.19%	71.89%	30.24%	-11.41%
Avg. ROI (per trade)	0.2363%	0.1863%	0.0863%	0.0363%	-0.0137%
No. of trades	833	833	833	833	833

Table 4. Performance with adjustment by the market (SPY certificate)

The numbers in table 4 show that hedging with the market results in a lower total ROI (~197%, when not considering transaction costs). In total our model required 833 trades over the entire period of time resulting in average ROIs of 0.24% (adjusted) and 0.49% (unadjusted) without considering transaction costs (0 bp). Average ROIs are calculated by dividing the total ROI by the number of trades. This results in an average ROI per trade. In Tables 4 and 5 we also display the resulting performances of our trading model for different trading costs expressed as basis points. Broker commission fees need to be subtracted, after all. This is generally a constant amount that highly depends on the broker. Usually, one can calculate it as $10 per transaction, independent of the traded volume. However, considering the return of our model it would be possible to subtract a constant broker fee and still receive a positive result.

Transaction Costs	0 bp	5 bp	15 bp	20 bp	25 bp
Total ROI	409.42%	367.77%	284.47%	242.82%	201.17%
Avg. ROI (per trade)	0.4915%	0.4415%	0.3415%	0.2915%	0.2415%
No. of trades	833	833	833	833	833

Table 5. Performance without adjustment by the market

Figure 1 shows the comparison of cumulated daily ROIs of our (unadjusted) trading model (without transaction costs) and cumulated daily price changes of the SPY certificate. We used the SPY certificate because it has very high trading liquidity and very accurately represents the current market value of the S&P 500 index. Thus, the SPY certificate is a suitable benchmark for our specific trading model. To compare its performance with our model the rule is as follows: buying the SPY certificate on the start day (price: 131.87) of the trading model and selling it on the last day (price: 124.99). The certificate lost 5.22% in this time; therefore, we outperformed it in almost every scenario from Tables 4 and 5 only when assuming transaction costs of 25bp and applying market adjustment that the certificate could beat our model.

It is obvious that a significant part of the performance of our trading model has been achieved during the period from the end of July until the beginning of September. This might be a result from a certain

market phase well suited for our model or statistical effects such as a temporary increase in correct signals for equities making large price changes during that time; more on that in the discussion section.

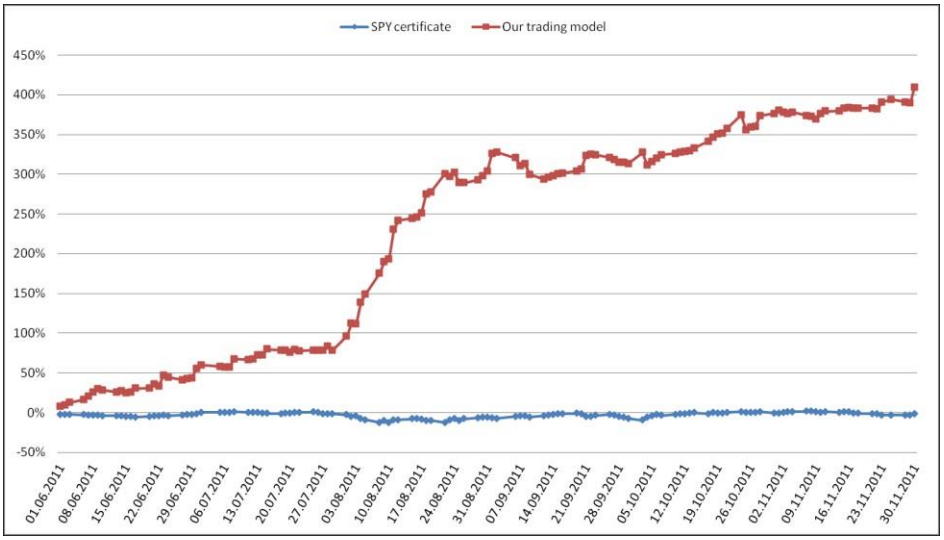

Figure 1. Comparison of cumulative returns on investments for our trading model and the SPY certificate from June 1 to Nov 30.

7. Discussion and Research Suggestions

Our findings confirm previous studies that investigated the relationship between online talk and financial markets and extend their validity to the case of multiple sources. It is obvious that predictive power rests in publicly available data. However, until now it has been uncertain whether this power has substantial value. Antweiler and Frank (2004) wrote: "This effect is statistically significant but economically quite small in comparison to plausible transaction cost." At least for the period between June 1 and November 30, 2011, we show that a positive ROI was achieved by trading based on public message board data – contrary to their results.

However, causality is uncertain and should be subject to further research. We would emphasize in particular that the period of time we used for our analysis is rather short. Thus, we cannot rule out that certain market conditions that were present during that period were responsible for the positive ROI of our model (e.g. the period from July to September when signals performed significantly better than on average). Long-term studies are required extending over different market and economic phases to address this limitation. Additionally, we used open and close prices, which is a simplification: realworld trading results will differ, either for better or worse.

Another shortcoming at this stage of our research is the comparison of models based on different data sources. We assume that Twitter, online message boards, and traditional news differ with respect to their predictive power for stock price movements. Thus, for further research we would like to explore differences in prediction ratios between the various sources we analyzed in this study. Additionally, we point to the method of sentiment analysis applied herein. It is clear that more sophisticated methods would provide better quality of correctly detected sentiment values in texts. Thus, the trading model's performance could potentially be improved by reaching higher quality in sentiment recognition.

Further it is necessary to analyze effects of applying different sentiment analysis methodologies. The quality a particular methodology delivers is probably related to the performance of the predictor. We hypothesize that a higher quality in sentiment analysis should lead to an increased number of correct predictions. Thus, as an extension of this work, we would like to compare the performance of the predictor when based on our own sentiment algorithms with the performance when other sentiment analysis tools are used for text classification. Improvement of performance could potentially also be achieved by applying separate sentiment analysis methods for different sources. E. g. Twitter

communication is different in nature from forum communication, thus an adapted classification method might improve sentiment quality (Sriram et al. (2010)).

References

Abbasi, A., Chen, H. and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Trans. Inf. Syst. (26:3) 1-34.

Anonymized (2012). Towards Universal Sentiment Analysis through Web Mining.

Antweiler, W. and Frank, M. (2004). Is all that talk just noise? The information content of Internet stock message boards. Journal of Finance. 59(3):1259-1295.

Asur, S. and Huberman, B.A. (2010). Predicting the Future with Social Media. arXiv:1003.5699v1.

Bagnoli, M. E., Beneish, M. D. and Watts, S. G. (1999). Whisper Forecasts of Quarterly Earnings per Share. http://ssrn.com/abstract=74369, retrieved 04/22/2012.

Berger, A.L., Della Pietra, S.A., and Della Pietra, V.J. (1996). A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics (22:1) 39-71.

Boiy, E. and Moens, M. (2009). A Machine Learning Approach to Sentiment Analysis. In Multilingual Web Texts, Information Retrieval (12:5) 526.

Bollen, J., Mao, H. and Zeng, X. (2010). Twitter mood predicts the stock market. arXiv Server.

Bollen, J., Pepe, A. and Mao, H. (2010). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. Proceedings from 19th International World Wide Web Conference Raleigh, North Carolina.

Brown, G.W. (1999). Volatility, sentiment, and noise traders. Financial Analyst Journal, Vol. 55 No. 2, pp. 82-90.

Brown, E.D. (2012). Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market. SAIS 2012 Proceedings. Paper 7.

Das, S. and Chen, M. (2007). Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53(9):1375-1388.

Das, S., Martinez-Jerez, A. and Tufano, P. (2005). eInformation: A Clinical Study of Investor Discussion and Sentiment. Financial Management 34(3):103-137.

Camerer, C. F. and Loewenstein, G. (2004). Behavioral Economics: Past, Present and Future. In Advances in Behavioral Economics, C.F. Camerer, G. Loewenstein and M. Rabin, eds. Princeton, NJ: Princeton University Press.

Choi, Y., Kim, Y. and Myaeng, S.H. (2009). Domain-Specific Sentiment Analysis Using Contextual Feature Generation. In Proceeding of the 1st international CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, ACM, Hong Kong, China, pp. 37-44.

Cohen, L. and Frazzini, A. (2008). Economic Links and Predictable Returns. Journal of Finance 63:1977-2011.

De Long, J., Shleifer, A., Summers, L. and Waldmann, R. (1990). Noise trader risk in financial markets. Journal of Political Economy 98:703-738.

Dion, M., Shaikh, V., Pessaris, A., Malin, S., Smith, R., Lakos-Bujas, D., Okamoto, S. and Hlavaty, B. (2011). News Analytics - Can They Add Value to Your Quant Process? – Using a Language Recognition Algorithm to Analyse News Flow. J.P. Morgan Europe Equity Research, June 3rd, 2011.

Fama, E. (1970). Efficient capital markets: a review of theory and empirical work. Journal of Finance. 25:383-417.

Fang F., Datta, A., and Dutta, K. (2012). A Hybrid Method for Cross-domain Sentiment Classification Using Multiple Sources. ICIS 2012, Orlando, Florida, USA.

Forster, M. R. (2002). Predictive Accuracy as an Achievable Goal of Science. Philosophy of Science (69:3), pp. S124-S134.

Gu, B., Konana, P., Rajagopalan, B. and Chen, M. (2007). Competition Among Virtual Communities and User Valuation: The Case of Investing-Related Communities. Information Systems Research 18(1):68-85.

Hey, T., Tansley, S. and Tolle, K. Eds. (2009). The Fourth Paradigm – Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington.

Hirshleifer, D. (2001). Investor Psychology and Asset Pricing. Journal of Finance 56(4):1533-1597.

Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. Joint 9th WEBKDD and 1st SNA-KDD Workshop. San Jose, CA.

Jones, C., (2006). A Nonlinear Factor Analysis of S&P 500 Index Option Returns. The Journal of Finance, Vol. LXI, No. 5, Oct. 2006, pp. 2325-2363.

Koch, S. and Schneider, G. (2002). Effort, co-operation and co-ordination in an open source software project: GNOME. Information Systems Journal 12(1): 27-42.

Krauss, J., Nann, S., and Schoder, D. (2012). Towards Universal Sentiment Analysis through Web Mining. Poster Session, European Conference on Information Systems, Barcelona, Spain.

Lin, C. and He, Y. (2009). Joint Sentiment/Topic Model for Sentiment Analysis. In Proceeding of the 18th ACM conference on Information and Knowledge Management, ACM, Hong Kong, China, pp. 375-384.

Lo, A.W. (2004). The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective, Journal of Portfolio Management: 30th Anniversary Issue), 15-29.

Mizumoto, K., Yanagimoto, H., and Yoshioka, M. (2012). Sentiment Analysis of Stock Market News with Semi-supervised Learning. Proceedings of the 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, 325-328.

Narayanan, R., Liu, B. and Choudhary, A. (2009). Sentiment analysis of conditional sentences. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, Association for Computational Linguistics, Singapore, pp. 180-189.

Neely, C. J., Weller, P. A. and Ulrich, J. M. (2009). The Adaptive Markets Hypothesis: Evidence from the Foreign Exchange Market. Journal of Financial and Quantitative Analysis (44:2)467-488.

Oh, C. and Sheng, O. (2011). Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. ICIS 2011, Shanghai, China.

Pang, B., Lee, L. and Vaithyanathan, S. (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, Association for Computational Linguistics, pp. 79-86.

Sabherwal, S., Sarkar, S. and Zhang, Y. (2008). Online talk: does it matter?. Managerial Finance 34(6):423-436.

Schoder, D., Gloor, P., and Metaxas, P. T. (2013). Social Media and Collective Intelligence – Ongoing and Future Research Streams. In Künstliche Intelligenz, Special Issue on Social Media. Forthcoming 2013.

Schumaker, R. and Chen, H. (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. ACM Transactions on Information Systems 27(2).

Schumaker, R., Zhang, Y., Huang, C.-N., and Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, Volume 53, Issue 3, June 2012, Pages 458–464.

Shmueli, G. and Koppius Otto R. (2011). Predictive Analytics. In Information Systems Research MIS Quarterly, Vol. 35, No. 3, pp. 553-572.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), Proceeding of the 33rd international ACM SIGIR

conference on research and development in information retrieval (SIGIR 2010), Geneva, Switzerland, July 19–23 (pp. 841–842). ACM.

Sprenger, T.O. and Welpe, I.M. (2010). Tweets and Trades: The Information Content of Stock Microblogs. http://ssrn.com/abstract=1702854, retrieved 04/22/2012.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62:1139-1168.

Tumarkin, R. and Whitelaw, R. (2001). News or noise? Internet postings and stock prices. Journal of Financial Analysts 57(3) 41-51.

Wang, S., Jank, W. and Shmueli, G. (2008). Explaining and Forecasting Online Auction Prices and Their Dynamics Using Functional Data Analysis. Journal of Business and Economic Statistics. (26:3), pp. 144-160.

Whitelaw, C., Garg, N. and Argamon, S. (2005). Using Appraisal Groups for Sentiment Analysis. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ACM, Bremen, Germany, pp. 625-631.

Xu, W., Li, T., Jiang, B., and Cheng, C. (2012). Web Mining For Financial Market Prediction Based On Online Sentiments. PACIS 2012 Proceedings. Paper 43.

Xu, S. and Zhang, X. (2009). How do social media shape the information environment in the financial market?. ICIS 2009.

Yerva, Surender Reddy, Miklós, Zoltán, and Aberer, Karl (2010). It was easy, when apples and blackberries were only fruits. EPFL working paper, http://infoscience.epfl.ch/record/151616/files/LSIR_WePS3_Paper.pdf, retrieved 03/30/2013.

Zhang, Y. and Swanson, P. (2010). Are day traders bias free? – Evidence from Internet stock message boards. Journal of Economics Finance 34:96-112.

Contribution V

Janetzko, D.; Krauss, J.; Nann, S. and Schoder, D. (2017) *"Breakdown: Predictive Values of tweets, Forums and News in EUR/USD Trading"* was published in the proceedings of the International Conference on Information Systems (ICIS) as a completed research paper

7-12-2017

Breakdown: Predictive Values of Tweets, Forums and News in EUR/USD Trading

Dietmar Janetzko
Cologne Business School, Cologne, Germany, dietmarja@gmail.com

Jonas Krauss
University of Cologne, Cologne, Germany, krauss@wim.uni-koeln.de

Stefan Nann
University of Cologne, Cologne, Germany, nann@wim.uni-koeln.de

Detlef Schoder
University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

Breakdown: Predictive Values of Tweets, Forums and News in EUR/USD Trading

Completed Research Paper

Dietmar Janetzko
Cologne Business School
Hardefuststr. 1, Cologne 50677, Germany
dietmarja@gmail.com

Jonas Krauss
University of Cologne
Pohligstr. 1, 50969 Cologne, Germany
krauss@wim.uni-koeln.de

Stefan Nann
University of Cologne
Pohligstr. 1, 50969 Cologne, Germany
nann@wim.uni-koeln.de

Detlef Schoder
University of Cologne
Pohligstr. 1, 50969 Cologne, Germany
schoder@wim.uni-koeln.de

Abstract

Predictive indicators for financial markets based on online buzz has been a frequent topic during the last years. Recent studies use a range of alternative sources for building these sentiment indices, with each purporting to have predictive value. Therefore a question mark remains regarding the comparability of findings across different types of sources, e.g. do indicators based on Tweets perform equally well or better than those built on news?

This study addresses how competing sentiment indicators affect EUR/USD trading. To identify the indicator having the best predictive value we estimate expected returns for individual sources and forecast models via backtesting. Our findings support the notion that the predictive value depends on the source of the sentiment-indicator, on timing aspects, with more recent sentiments having greater predictive strength, and on the type of rule (e.g. buy / sell) harnessed.

Keywords: Twitter, sentiment analysis, online communities, predictive modeling, financial prediction, natural language processing, data analysis

Introduction

Signal or noise? This is the classical question that guides a considerable amount of work in computational finance and in data mining. It is also the unifying theme of studies that examine whether sentiment indicators help to forecast volatility on global financial markets. The possible influence of sentiments on financial markets had already caught the interest of John Maynard Keynes (1935). However, empirical tests of this intuition became only possible at the beginning of the second millennium, when Web 2.0 applications like message boards and social media, in particular Twitter, facilitated large-scale collection of sentiment-related data suitable for financial modeling. But not only technical developments pave the way to using sentiments expressed on social media. In April 2013, the US Securities and Exchange Commission (the US market-regulator) gave social media outlets the go-ahead to publish corporate information (US Securities and Exchange Commission, 2013). This legal decision is an accolade for social media as a channel for business communication. It is likely to contribute further to discussions on social media, rich in sentiments.

Prediction of stock movements is the focus of the majority of work that deployed sentiment indicators as explanatory variables. An early study of Tumarkin and Whitelaw (2001) found little or no evidence

that sentiments expressed on Internet message boards predicted industry adjusted stock returns. On the contrary, their findings suggested that developments on the stock market prompted discussions on message boards. Their results in 2001 were far from unequivocal and in fact more recent work finds some evidence that sentiment content sourced from the Internet may predict the ups and downs on stock markets. Similar to the work of Tumarkin and Whitelaw (2001), Antweiler and Frank (2004) investigated whether sentiments on Internet message boards predict stock returns. In their study, sentiments turned out to have a small, but negligible predictive effect. Bollen et al. (2010) collected and classified Tweets to forecast daily closing values of the Dow Jones Industrial Average. According to the authors, the Tweets reached an accuracy of 86.7% when predicting the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%. Work by Nann et al. (2013) extended the singlesourced prediction models, by aggregation of sentiment indicators from multiple online sources. Using this approach the out-of-sample prediction accuracy turned out to range between 53% and 67%.

On the whole, predictive modeling via sentiment-related data from finance microblogging has come a long way. While early research found that discussions followed market movements, the majority of more recent studies detect predictive value in online data with an above-chance level for many features of financial markets like trading volume, volatility and return. However, the field as a whole is still in its infancy because a number of open questions remain. A particularly pertinent one is the robustness of the findings which obviously is essential for the field. For instance, Sprenger et al. (2014) rightly point out that the sample period in the widely cited study of Antweiler and Frank (2004) is the year 2000. This year was characterized by unrealistic expectations and discussions associated with the dotcom bubble, which puts a question mark over the robustness of the findings. Moreover, predictive effects found in previous studies were often economically small to moderate (Ranco et al., 2015). Other studies have not found predictive effects at all, e.g. Oliveira et al. (2013). The vast majority of studies use end-of-day stock prices for estimating returns, which limits the ability to simulate actual trading thus obstructing robustness even further.

Another open issue relates to the choice of the data source: does the predictive value of a sentiment indicator depend on the online source it is built on or do different sources deliver the same quality of prediction? And, if quality varies, which online source is optimal? For instance, Antweiler and Frank (2004) built their sentiment indicator around messages from online forums Yahoo! Finance and Raging Bull, blogs were the subject of analysis in Melville et al. (2009), Bollen et al. (2010) used communication on Twitter, Oh and Sheng (2011) chose StockTwits as their source and Chen et al. (2014) looked at the popular investor blog Seeking Alpha. This limits the ability to objectively compare findings from each study. It remains unclear whether their findings differ because of different data sources, different methods of sentiment analysis, observed market phase or different time units the data was collected for.

This paper has been motivated by the need for comparability between sentiment sources and by questions of robustness and varying predictive value across different sources. As the majority of previous research is concerned with stocks and indices (see next section) this paper also contributes by studying sentimentbased predictions for another kind of financial instrument: currencies. Comparability is considered by examining a number of social media data sources and studying their predictive performance in a unifying framework. Robustness will be addressed by a sampling period of more than 40 months, which is considerably longer than those used in previous studies. Further, our sample consists of intraday data allowing for a more realistic and more significant trading simulation. The sample will be analyzed via a backtesting approach designed to identify parameters that increase cumulative returns. The assessment of individual predictive values from different online sources is addressed by taking data samples from multiple popular social media sources and applying the same backtesting approach to each of them. This will enable the objective comparison of results across sources.

This study is organized as follows: Firstly, the literature review puts our own work into the context of research concerned with the predictive power of online data, in particular sentiments for global financial markets. Secondly, we will describe our data sample with regard to its origin and pre-processing steps taken. Thirdly, the backtesting strategy and trading rules will be described. Fourthly, results of the analysis are presented. Finally, findings of this study are discussed and suggestions for future research are given.

Previous Work

During the first years of the second millennium, the volume of user-generated content on social media and thus social media data was comparatively small. Communication was scattered across a number of blogs as today's social media giants, in particular Twitter, had not yet been founded. Still, early on some researchers explored the possibilities to derive predictions from social data. The work by Bagnoli et al. (1999) is a precursor of these types of studies. Using a variety of information from online and offline sources to derive forecasts of quarterly stock returns, their results outperformed competitive forecasts of individual analysts assembled by the First Call Corp. (first call forecasts). To the best of our knowledge, the first publication concerning the use of social media data for financial forecasting was a study by Tumarkin and Whitelaw (2001). The authors examined a possible link between sentiments and posting volume recorded on a major financial message board, on the one hand and stock returns and trading volume on the other. There were some hints that message board activity and stock returns were related on days of abnormal message board activity. But in this early study neither message board activity nor sentiments facilitated predictions of future returns or volume. On the contrary, the findings suggested that developments on the stock market prompted discussions on message boards. Similar to this work, Antweiler and Frank (2004) investigated whether Internet message board activity or sentiment is predictive for stock returns or market volatility. They found that board messages help to predict volatility on stock markets. The effect of board messages on stock returns was statistically significant, but economically small. Other studies of the first decade of the second millennium used data collected on blogs to determine correlations with stock market movements (De Choudhury et al., 2008). The findings discussed so far were corroborated and extended by a study of Sabherwal et al. (2008) which also found that online talk has the power to predict next-day stock returns. To sum up, studies till 2008 mainly relied on data from financial blogs or Internet message boards and usually reported none or small predictive effects.

Twitter went live in July 2006. Due to its enormous popularity and its relatively simple API the attention of many researchers interested in social media data shifted from blogs to Twitter. A few years later, the data made available via Twitter prompted a string of publications. While Oh and Sheng (2011) looked at a comparably small subset of messages from StockTwits, Bollen et al. (2010) collected a large amount of 9.85 million postings from Twitter. The work by Bollen et al. (2010) suggests that some public mood states expressed on Twitter are predictive of changes in Dow Jones Industrial Average (DJIA) closing values. Only a comparatively small number of more recent studies investigated the characteristics of the information sourced from financial microblogs that translate into predictive power for movement on financial markets. Work by Sprenger et al. (2014) revealed that stock microblogs may contain alpha information, i.e. information that is not yet reflected in price levels on asset markets, thus leading to future movement in line with the efficient market hypothesis (Fama, 1970). In addition, microblogging forums efficiently aggregate this information via mechanisms like retweets and follower relationships. A study by Rao and Srivastava (2014) looked into the timing of this information and suggests that movements on financial markets are greatly affected in the short term by discussions on financial microblogs. Recent work by Veiga et al. (2016) analyzes the different impacts of traditional news and social media on stock market volatility and trading volume. Their findings show that both sources affect markets in different ways.

Compared to stock market predictions there is only a relatively small number of studies that make use of sentiment-based indicators to predict exchange rate movements. Work by Papaioannou et al. (2013) and Janetzko (2014) on predictive modeling of the EUR/USD change rate and a study of Crone and Koeppel (2014) that harnessed sentiment indicators as explanatory variables to predict the AUD/USD currency pair, provided evidence that sentiment indicators have the potential for correct predictions in more than 50% in out-of-sample tests. In several regards, our work goes beyond these three studies. Firstly, we extend the data base studied from single to multi source (compared to Papaioannou et al. 2013 and Janetzko 2014, both only uitilizing Twitter as source). Secondly, we set the focus on intraday data. Finally, we compare predictive values across the different data sources (compared to Crone and Koeppel 2014) with an emphasis on robustness.

Data

We collected online messages about currencies Euro (EUR) and US Dollar (USD). The messages consisted of Tweets, forum posts and news gathered between June 19, 2012 and November 12, 2015. Sentiment values from these messages were aggregated using units (time slices) of 10 minutes. Prior literature uses either end-of-day EUR/USD price data, e.g. Crone and Koeppel (2014), or hourly data,

e.g. Papaioannou et al. (2013), to build prediction models. To go a step further we chose to use a shorter time slice of 10 minutes. We assume that an aggregation of sentiment values on smaller time units is able to deliver more information as opposed to aggregating longer time units. Forex markets have an average daily turnover estimated at around USD 5.1 trillion per day and the exchange rate EUR/USD is one of the most traded pairs (BIS, 2016). Thus, choosing a 10 minute time slice as the basic unit of data-recording seems to be a reasonable trade-off between the frequency of online messages and trading volume. The different data sources we considered differ with regard to the frequency of signals, e.g. Tweets recorded on the time axis (see next section). Data-recording units of 10 minutes allowed data-analysis units (prediction base) of different spans, e.g. 20 minutes, and thus facilitates testing our research questions.

Origins of the data were major financial microblogs and other type of online sources. Message types can be characterized as follows:

- *Tweets* are short messages (140 characters maximum length) collected from Twitter and StockTwits. Usually Tweets appear in higher frequency.
- *Forum posts* collected from Forexfactory are typically longer than Tweets and frequently follow a dialogue pattern.
- *News* are traditional (online) news articles. Typically, news articles have longer content and appear less frequent (compared to Tweets or forum posts).

Our analysis was carried out with the goal of finding the most valuable type of source for predicting future exchange rates of EUR/USD. Selecting appropriate messages and pre-processing them involved 4 steps:

1. *Identification of communication about EUR and USD*. When harvesting forum posts from online message boards, we considered messages in specific sub-forums where EUR and USD are discussed. In case of Tweets we made use of Twitter's cash tag system (*$EURUSD*). For news articles and forum posts we used a keyword-based Naïve Bayes classifier that harnessed concepts like *Euro*, *Dollar* and *ECB* (European Central Bank) to identify whether or not a news article is concerned with the subject. For a complete list of terms see the appendix.
2. *Identification and filtering of spam*. Our spam filter searched for insulting communication ("rants" or "flames") devoid of relevant information. Most of these posts can be identified by their scurrile or nasty language. As in the first step, we applied Naïve Bayes classification to accomplish this task.
3. *Identification of negations*. Appropriate processing of negations is essential to distinguish between messages that mention expressions like *buy* as opposed to *don't buy*. To achieve this objective we used part-of-speech tagging. Via a moving window technique applied to all words of a message we analyzed whether the word in focus was a sentiment-related concept. If such a concept was found, its vicinity was scanned for negation terms like *not* or *don't*. In case such a term was found, the whole phrase was substituted by an expression of the bag-of-words from the opposite sentiment class, e.g. *don't buy* was replaced by *sell*.
4. *Sentiment scoring of messages*. The final step led to a sentiment score for each message, again by using a Naïve Bayes classifier as it is widely applied in research (Medhat et al., 2014). The quality of the classifier depends on how well it is adapted to the analyzed content. This is often characterized by an author's individual choice of words when expressing a positive or negative opinion (Nann et al., 2013). To address this problem, we assembled a sample of approximately seven hundred Tweets and posts from the available dataset and manually labeled each message as expressing positive or negative sentiment. Lists of positive (e.g. buy, long, call) and negative words (e.g. sell, short, put) were curated manually to create the bag-of-words deployed for classification. A Naïve Bayes classifier was used to calculate the ratio of occurrences of positive and negative words in a forum post, Tweet or news article. If the number of positive was larger than the one for negative words, a message was classified as positive and vice versa. If the ratio was not decisively larger for one of the two classes, a message was classified as neutral. We opted

for a simple tertiary setup of the sentiment variable as it corresponds well to the types of signals we want to derive at a later stage: "buy" and "sell". Figure 1 illustrates the four steps of our sentiment detection process.

Figure 1. Selecting and Pre-Processing Messages

To estimate the classification performance of our classifier we let it calculate the sentiment value for each message from our manually labeled sample, compared results with manually tagged values and calculated precision and recall ratios. With a precision of 0.83 and a recall of 1 (every message was classified) the classifier achieves a quality measure F of ~0.907. Message counts for each sentiment class (positive, negative and neutral) were calculated for each time slice, which led to three values (pos, neg, neu) for each source per time slice. The resulting time series were deployed in all analyses described below.

It is our main hypothesis that for different types of sources the individual predictive value varies. For instance, it seems plausible to assume that Tweets have a lower half-life period than news articles which could make them better suited for short-term predictions but less for long-term developments. We can argue that because positive messages occur more frequently (a phenomenon commonly found in literature concerned with detecting sentiment in Tweets, e.g. Sprenger et al. 2014) considering negative messages only have a higher potential of indicating exchange rate movements.

All types of messages were collected via public APIs or crawling techniques. All sources, their types and number of collected messages are given in Table 1.

Source	Type	# Messages
pos.www.forexfactory.com	Positive Forum Posts	131,716
neg.www.forexfactory.com	Negative Forum Posts	81,319
neu.www.forexfactory.com	Neutral Forum Posts	201,681
pos.stocktwits.com	Positive Tweets	35,870
neg.stocktwits.com	Negative Tweets	28,230
neu.stocktwits.com	Neutral Tweets	72,419
pos.twitter.com	Positive Tweets	5,060,738
neg.twitter.com	Negative Tweets	3,334,132
neu.twitter.com	Neutral Tweets	2,586,531
pos.www.marketwatch.com	Positive News	5,483
neg.www.marketwatch.com	Negative News	3,179
neu.www.marketwatch.com	Neutral News	3,776

Table 1. Sources, Types and Number of Messages

Method

Backtesting

Backtesting is a simulation method that makes use of historical data to estimate profits and losses, which an investment strategy could have earned in the past. While backtesting is a widely used method for evaluating investment strategies, its results do not necessarily generalize to future situations. Backtesting is often criticized as it is prone to overfitting and data snooping across different model variations, which often remain unreported (e.g. Bailey et al., 2014). The motivation to harness backtesting in this paper is to generate hypotheses on parameter settings, and hence our backtesting

strategy is explorative. We backtest each set of sentiment data with a number of model variations and rank results by cumulative returns.

Formal Setup

Backtesting is based on sentiment information collected from social media between June 19, 2012 midnight and November 12, 2015, 11:30am. Its purpose is to generate buy or sell signals, in order to maximize returns of a simple cash portfolio by EUR/USD trading. The setup used for backtesting a range of different sentiment indices can be more formally described as follows.

A fixed-size moving window technique with window-size m was applied to the sentiment indices to generate trading signals. The median has been preferred over the mean because the median is known to be more robust than the mean vis-à-vis outliers. But clearly, when m=2 the moving median becomes the moving average. We will use S_{jt} to refer to $j = \{1 \ldots J\}$ sentiment indices evaluated regarding their efficiency when used in EUR/USD trading. Accordingly, the sentiment score of index j measured at time t is expressed as $S_{jt.}$ Let

$$E_{kt,S_{jt}}$$

denote one of $k = \{1 \ldots K\}$ binary trading rules that deploy sentiment index j at time t with $t = \{1 \ldots T\}$ to generate a buy (1) or sell (-1) signal on the basis of sentiment information at time t-m for a trading transaction at time t. In our study, we considered 12 different financial microblogs or similar sources to construct sentiment indices. For control reasons, we also assembled a sequence of scores by randomly reshuffling sentiment scores from all 12 sentiments indices of length T. This led to J=12+1 sentimentbased indices. In what follows, we simply equate T with the total number of 10-minutes units considered. Each trading rule $E(t_k, S_{jt})$ under consideration is a three-place function that maps into {1, −1}. The central element of each trading rule is a simple comparison of $S_{jt.}$ with a moving median MM_{tSj-m} which can be defined as follows

$$E_{kt,S_{jt}} = \begin{cases} -1, & MM_{t-m}^{S_j} > S_{jt} \\ 1, & MM_{t-m}^{S_j} \leq S_{jt}. \end{cases}$$

This means that trading using K=2 trading rules follows a counter-cyclical trading scheme. A short signal (-1) was generated, whenever the sentiment value S for a given sentiments index j generated by the moving median MM at time t-m was greater than the same sentiment index at time t. Vice versa, a long signal was produced, whenever the sentiment value S for a given sentiment index j generated by the moving median MM a time t-m was smaller than the same sentiment index at time t. The trading rules are not triggered by the magnitude of sentiments per se but by a comparison of a rolling sentiment $S_{jt.}$ against a rolling sentiment reference MM_{tSj-m}. This means that the trading rules constructed and tested in this paper assume a functional relationship between conditional sentiments and returns. However, the relationship between sentiments and returns is considered to be indirect in that not the sentiment magnitude per se was the basis for buy/sell but its comparison with the sentiment calculated via moving median MM at time t-m. The application of the K=2 trading rules to J=13 sentiment indices generates J binary vectors of trade signals. Under the EUR/USD backtest trading scheme, a long position (buy) was encoded by 1 and a short position was encoded by -1. For all units or time slices T and for all sentiments vectors J, cumulative returns were calculated by multiplying the vector of the differenced close EUR/USD exchange rate (rate of change) during the time under consideration by the vector of trade signals. This was followed by cumulatively summing up returns obtained from one time slice to the following one. For instance, if from one time slice t to the following time slice $t+1$ the EUR/USD exchange rate increased by x cent and the trade signal was 1 for this time slice $t+1$, then the magnitude of the increase x qualified as a positive return. Using just the close of the EUR/USD exchange rate for backtesting amounts to a cash portfolio of with 1 Euro. But clearly the percentage of cumulative returns can be calculated with such a portfolio and the results apply to large cash portfolios as well.

Analysis and Results

Descriptive Statistics

The descriptive statistics of sentiment indices prior to the analysis is presented in Table 2. Results indicate that most indices exhibit a high degree of variability in themselves and among each other. On the one end of the spectrum, sentiment indices gained from *pos.twitter.com* and *neg.twitter.com* score high on measures of dispersion. The same kind of measure is comparatively low when looking at sentiment indices sourced from *neg.stocktwits.com* and *pos.stocktwits.com*. The order of integration of all sentiments was found to be 0 for all sentiment indices and 1 for the EUR/USD exchange rates, as evidenced by the augmented Dickey-Fuller unit root test (e.g. Banerjee et al., 1993) for $p < 0.01$. These results indicate that sentiment indices are stationary while the EUR/USD exchange rate is not. Though it is well known that the ADF test has a low power, the number or time units under consideration T of each time series is very high. We can assume that the standard error decreases with rising T and that the power is expected to increase. Therefore, the results indicate a mismatch of the order of integration between all sentiment indices on the one hand and the EUR/USD exchange rate on the other. This finding suggests that standard tests of cointegration are not applicable and the same applies to Granger causality testing. Sentiment indices obtained from the "news" class (*pos.www.marketwatch.com*, *neg.www.marketwatch.com*, *neu.www.marketwatch.com*) do not appear in Table 2 since they all showed a low degree of variability. The reason for this is their characteristically less frequent appearances, compared to Tweets or forum posts. For this reason we did not continue to use this source for our analysis.

Sentiment Index	Min	Max	Median	Mean	Sum	Variance	StdDev	OI	p/ADF
neu.www.forexfactory.com	0	32	0	1.160	144759	3.714	01.92	0	.01
neg.www.forexfactory.com	0	13	0	0.470	58728	0.866	00.93	0	.01
pos.www.forexfactory.com	0	19	0	0.750	93976	1.694	01.30	0	.01
neu.twitter.com	0	963	9	15.390	1911267	442.453	21.03	0	.01
neg.twitter.com	0	814	4	20.040	2488928	1306.452	36.14	0	.01
pos.twitter.com	0	1207	6	29.730	3692370	2619.452	51.18	0	.01
neu.stocktwits.com	0	25	0	0.410	52141	0.792	00.89	0	.01
neg.stocktwits.com	0	24	0	0.160	20348	0.258	00.50	0	.01
pos.stocktwits.com	0	19	0	0.200	25732	0.348	00.59	0	.01
Random	0	899	0	8.600	1067788	703.635	26.52	0	.01
EUR/USD	1.205	1.390	1.322	1.320	163909	0.001	00.04	1	0.01

Table 2. Descriptive Statistics of Sentiment Indices (T=124157)

Cumulative Returns

The forex market, especially the currency pair EUR/USD, is known to be very efficient in the way that it quickly integrates new market-relevant alpha information into exchange rates. Work by Sprenger et al. (2014) suggests that sentiments gathered in financial microblogs may provide alpha information, which could facilitate temporary above-market returns. Design implications are obvious. A small time window for using this information rules out long-memory models and techniques, e.g. a recursive time window. Instead, modeling should focus on a moving window that has either no memory or a short memory that is small enough to capture alpha information not yet incorporated into exchange rates. Results of our backtesting analysis with a time window of m=2 time slices of 10 minutes, over the sampling period (June 19, 2012 midnight and November 12, 2015, 11:30am), are presented in Figure 2.

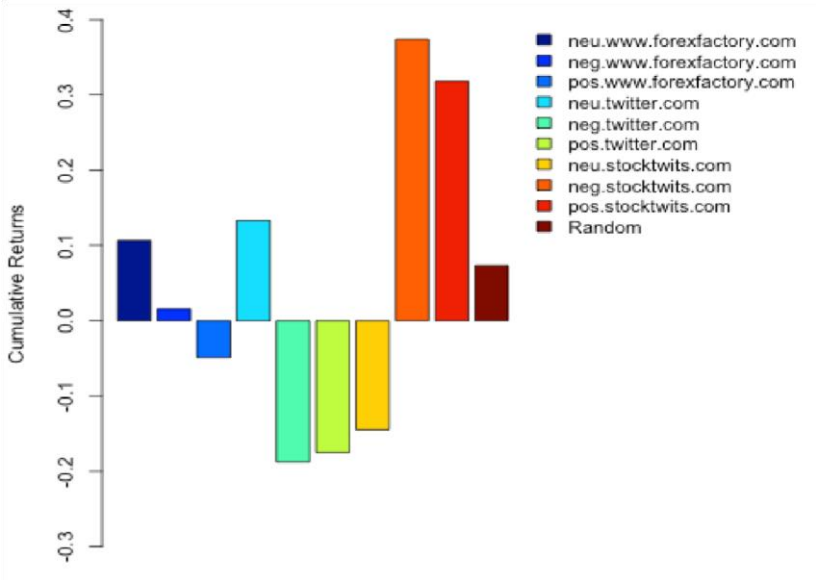

Figure 2. Cumulative Returns by Sentiment Index (longitudinal view)

Figure 2 shows that the majority of sentiment indices, when used to inform EUR/USD trading with the parameters mentioned above, facilitate a development of returns that is characterized by high volatility and mean reversion. This performance pattern is typical of the random index. In this regard, the majority of sentiment indices exhibits a performance pattern that does not differ from that of the random index. There are, however, some sentiment indices that deviate from this pattern. In particular returns generated by sentiment indices from *neg.stocktwits.com* and *pos.stocktwits.com*, often co-move and show a comoving upward trend. Figure 3 shows the cumulative returns as they develop on the time axis.

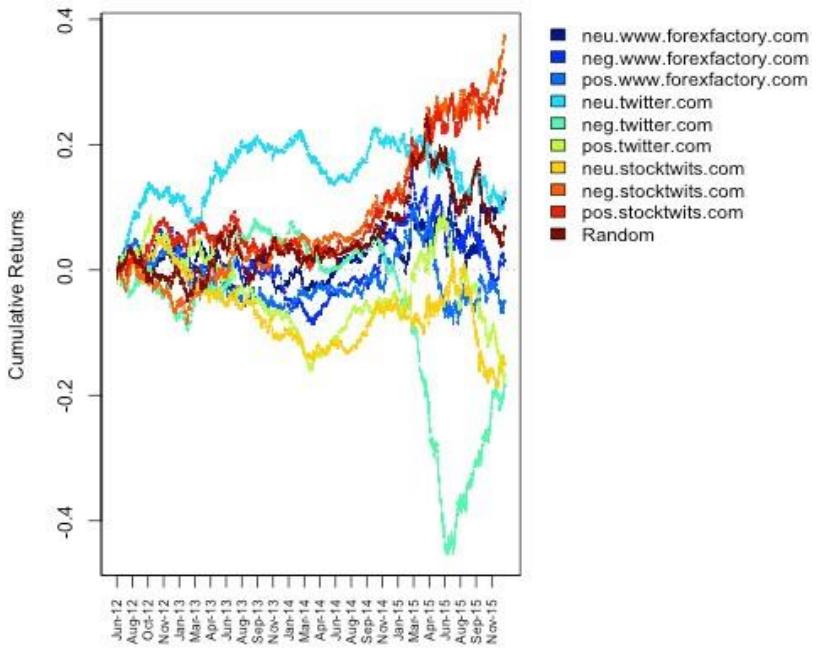

Figure 3. Cumulative Returns by Sentiment Index (cross-sectional view)

It shows that *neg.stocktwits.com* and *pos.stocktwits.com* are sources of sentiment data that provided the best performing sentiment indices. Though, one might argue that indices leading to negative returns could possibly be adjusted by simple reversals of comparisons in the algorithm, we present these results as they occurred. Table 3 shows cumulative and abnormal returns for sentiment indices and EUR/USD.

Sentiment Index	Cumulative Return	Cumulative Abnormal Return
neu.www.forexfactory.com	10.65%	25.37%

neg.www.forexfactory.com	1.57%	16.29%
pos.www.forexfactory.com	-4.86%	9.90%
neu.twitter.com	13.33%	28.03%
neg.twitter.com	-18,73%	-4.03%
pos.twitter.com	-17,48%	-2.71%
neu.stocktwits.com	-14,46%	0.32%
neg.stocktwits.com	37.39%	52.17%
pos.stocktwits.com	31.82%	46.59%
Random	16,51%	28.16%
EUR/USD	-14.74%	0.00%

Table 3. Cumulative and Abnormal Returns by Sentiment Index

Differential Effects of Buy and Sell Rules Informed by Sentiments

Why do some sentiment indices perform well while others do not? In the preceding sections we mainly looked at the outcome of sentiment-informed trading without investigating the conditions that led to the observed effects. An obvious starting point to examine not only the effects but also its underlying conditions is to split up the overall effects found for each sentiment index with regard to the two trading rules used. Both the buy and the sell rule follow a counter-cyclical pattern. The buy rule is triggered whenever the sentiment is relatively low, and the sell rule applies when the sentiment is relatively high.

Returns	Overall	Buy/ Sum	Sell/ Sum	Buy/ Mean	Sell/ Mean	Skewness Buy & Sell	Skewness Buy	Skewness Sell	Sharpe Ratio
neu.www.forexfactory.com	0.1065	-0.0274	0.1339	-8.68E-07	1.45E-06	17.270	-51.974	-54.636	0.0290
neg.www.forexfactory.com	0.0157	-0.0728	0.0885	-3.15E-06	8.75E-07	35.380	-27.467	-65.838	0.0040
pos.www.forexfactory.com	-0.0486	-0.1049	0.0564	-3.66E-06	5.90E-07	47.540	-11.115	-71.691	-0.0130
neu.twitter.com	0.1333	-0.0140	0.1473	-2.47E-07	2.18E-06	21.890	-41.541	-61.718	0.0360
neg.twitter.com	-0.1873	-0.1743	0.0130	-3.56E-06	-1.73E-07	-39.090	-93.357	-14.781	-0.0500
pos.twitter.com	-0.1748	-0.1680	0.0067	-3.23E-06	-9.34E-08	0.9240	-51.520	-56.083	-0.0470
neu.stocktwits.com	-0.1446	-0.1530	0.0083	-6.57E-06	8.26E-08	-22.910	-108.970	-23.306	-0.0390
neg.stocktwits.com	0.3739	0.1063	0.2676	8.69E-06	2.39E-06	55.370	0.3750	-64.899	0.1000
pos.stocktwits.com	0.3182	0.0785	0.2398	5.58E-06	2.18E-06	59.060	11.541	-70.789	0.0850
Random	0.1651	0.0019	0.1632	8.69E-06	2.39E-06	43.820	-82.856	-33.715	0.0440

Table 4. Overall and Partial Effect of Sentiment Indices (Percentage of Cum. Returns)

The results presented in Table 4 illustrate the overall effect (buy & sell) for each rule along with a breakdown into the partial effects of the buy rule and the sell rule, respectively. A closer look at the breakdown reveals that sell rules often seem to perform better than buy rules.

Distributions of returns reveal that there are differences in the skewness between sentiment indices and the combined rule (buy & sell), the sell and the buy rules. For instance, with regard to the worst trading rule, *Random*, there is an overall effect of 1% cumulative returns, while the best trading rule, *neg.stocktwits.com,* has an overall effect of 37.39% cumulative returns. Further examination of the contribution of buy and sell rules applied to *pos.stocktwits.com,* shows that the buy rule contributes only 10%, while the sell rule contributes 27% to the overall returns. A similar pattern can be observed in all sentiment indices with positive overall returns (*neu.www.forexfactory.com, neg.www.forexfactory.com, neu.twitter.com, neg.stocktwits.com, pos.stocktwits.com, Random*).

What is the reason behind the higher overall performance of the sell rule? Considering that the exchange rate of EUR/USD drastically fell during the analyzed period of time, a plausible explanation seems to be that under conditions of falling prices, selling is more profitable than buying. However, this conjecture

is at odds with findings obtained for the *Random* index. Here, the sell rule is only marginally better than the buy rule. But this does not explain the large gains observed in other cases where the sell rule is applied (*neg.stocktwits.com* and *pos.stocktwits.com*). Obviously, selling at random points in time during a phase of declining prices leads to small, but negligible cumulative returns as evidenced by the performance of the *Random* index. The simple strategy to sell at random points is outperformed by strategies that apparently have a higher likelihood to identify the right moment for a sell transaction.

Figure 4 illustrates the effect of the sell rule via a histogram of the returns (buy & sell, buy, sell) of the best scoring sentiment index *neg.stocktwits.com*. Along with the added skewness values the figure confirms that the overall effect of the sell rule on returns is higher than that of the buy rule. This is echoed by higher skewness values of the returns gained with the sell rule when compared to the gains of the buy rule.

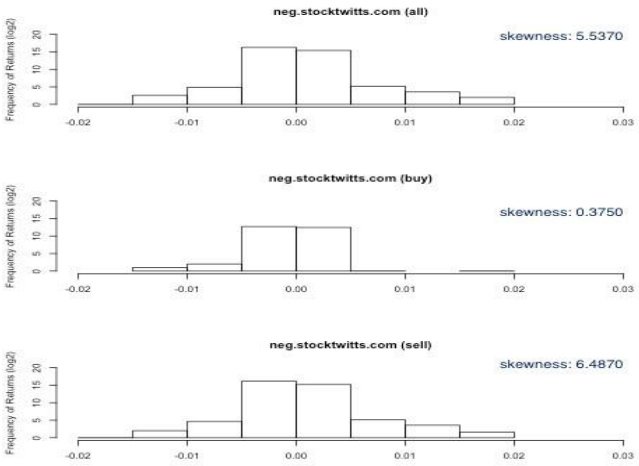

Figure 4. Histogram of the Returns for the Best Performing Sentiment Index

It has been mentioned that the buy rule contributes only about a third to the overall effect of high performing sentiment indices. But this finding reflects only an overall view as it does not consider potential differences in the average return of buy and sell rules. In fact, the average performance of the buy rule (buy/mean) is higher than that of a sell rule (sell/mean). Looking at the average performance reveals also that *neg.stocktwits.com* and *pos.stocktwits.com* stand out from all other sentiment indices because their performance is consistently good for both buy and sell rules. All other sentiment indices show a poor average performance either in their buy or sell rules or even in both of them.

Discussion

By definition, efficient markets swiftly incorporate new market-relevant information into asset prices so that the latter "fully reflect" the former (Fama, 1970). This more general rule also applies to marketrelevant information obtained from financial online sources. This is the most plausible reason why cumulative returns turned out to be highest when a small prediction window *m* in combination with a fixed-size moving window were chosen. Both are geared towards a short memory process where mainly recent events have an effect on the movement of the market. The forex market is characterized by high trading volume and high volatility. Unlike stock trading, it is not bound to a stock exchange, but currencies are traded around the clock, regardless of market hours. Our dataset fits these characteristics by having 10-minute units around the clock.

The main goal of this study was an estimation of the predictive value of different source types for purposes of trading in forex markets. As corollary of our findings we could highlight the conditions that actually led to higher performance. Comparing our results with a random index showed that the observed performance cannot be explained entirely by general market developments. Thus we have a strong indication for actual predictive value in the analyzed data and can rule out that positive returns were completely caused by external influences like a bear market in EUR/USD. Our results show a clear winner: sentiment messages from StockTwits. We can state that Tweets clearly yield superior forecasting power, compared to traditional news and forum posts – at least in the case of forex markets and a small

prediction window. By analyzing buy and sell signals separately we found that sell signals seem to have a higher predictive value for EUR/USD. In addition to these findings we were able to demonstrate that methods of forecasting financial developments in stock markets established in previous research can successfully be applied to currency trading as well. Future research is necessary to compare the performance of Tweets, forum posts and news on longer prediction windows keeping in mind that eventually predictive value will dissipate completely because markets swiftly incorporate new information into their prices (especially in the case of forex).

Since our goal was not to achieve a new state-of-the-art return on investment by using sentiment indicators, but rather to estimate the predictive value of different kinds of sentiment sources, we proceed with discussing the key characteristics of the sources used. The best sentiment indices in our sample were gathered from *pos.stocktwits.com* and *neg.stocktwits.com*. Both fall into the source class "tweets". They clearly outperform twitter.com series from the same source class. While Twitter is a micro-blogging platform not concerned with a special topic (yet carrying a huge amount of communication concerned with financial markets), StockTwits is solely focused on discussing trading and financial markets. Therefore it is plausible to argue that the outperformance of its results can be rationalized with the specialisation of its community. StockTwits contributors are "individual, institutional and hedge fund managers" as stated by Pierce Crosby of Quantopian in a blog post (Crosby 2015). Their expertise in trading could explain why the predictive value of this source ranks highest.

Having a closer look at the results, we observe that messages which carry sentiment (positive or negative) on StockTwits outperform the neutral indicator from the same source. It is reasonable to argue that neutral messages have lesser predictive value so this can be considered an expected result. For Twitter, however, neutral messages rank highest (compared to messages on Twitter which carry sentiment). Here we can only assume that the lesser degree of financial expertise of its users leads to a decreased usefulness in forecasting. For example positive Tweets from Twitter (*pos.twitter.com*) even have a negative relationship, which indicates that this information is already reflected into EUR/USD prices. This might be due to the fact that there are numerous systems in place which collect tweets from Twitter and base trading decisions on it.

Other kinds of sources could not achieve equal predictive value. Both other classes, “news” and “forum posts” represented by MarketWatch and Forexfactory respectively, are characterized by longer textual content and less frequency, in comparison with Tweets. Considering characteristics of the forex market, it could be reasoned that a lower number of new messages per time has a strong negative impact on the predictive value.

The main contribution of our paper is twofold: By setting up an analytical framework that allows to compare the effects of a range of different indices we were able to show that some sentiment indices clearly outperformed others. Going beyond a pure effect analysis we could (i) progressively narrow down conditions that lead to increased performance which led us to uncover a dual effect of buy and sell trading rules and (ii) rule out plausible alternative explanations (general market development). Splitting up the overall effect of cumulative returns it became obvious that sell rules account for about two thirds while buy rules account for only one third of the overall effect. In contrast to other measures, e.g. skewness, returns are an extensive quantity that obeys the rule of additivity, and hence a splitting of the overall effects of cumulative returns is possible. Partial return effects were jointly generated by a frequent application of a comparatively moderately effective rule (sell) and rare application of a relatively effective rule (buy).

Our models were built for trading the exchange rate EUR/USD. Future research is required to investigate whether our finding regarding differential effects hold true for other exchange pairs as well and whether the findings of this research also apply to stocks and commodities trading. Based on preliminary analysis for other asset classes we assume that our findings can be generalized. More work and further research will be necessary to answer this question more profoundly.

References

Antweiler, W., and Frank, M. Z. 2004. “Is all that talk just noise? the information content of internet stock message boards,” *The Journal of Finance* (59:3), pp.1259–1294.

Bagnoli, M., Beneish, M. D., and Watts, S. G. 1999. “Whisper forecasts of quarterly earnings per share,” *Journal of Accounting and Economics* (28:1), pp. 27-50.

Bollen, J., Mao, H., and Zeng, X. 2011. "Twitter mood predicts the stock market," *Journal of Computational Science* (2:1), pp. 1-8.

Bailey, D. H., Borwein, J., de Prado, M. L., and Zhu, Q. J. 2016. "The probability of backtest overfitting," *Journal of Computational Finance.* (20:4) pp. 39-69.

Banerjee, A. Dolado, J. J., Galbraith, J. W. and Hendry, D. F. 1993. "Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data," *Oxford University Press*, Oxford.

BIS 2016. "Triennial Central Bank Survey of foreign exchange and OTC derivatives markets in 2016," *Bank of International Settlements* (BIS), Online Available at https://www.bis.org/publ/rpfx16.htm?m=6|35.

Chen, H., De, P., Hu, Y., & Hwang, B. H. 2014. "Wisdom of crowds: The value of stock opinions transmitted through social media," *The Review of Financial Studies*, 27(5), pp. 1367-1403.

Crone, S. F., and Koeppel, C. 2014. "Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons," in *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2104 IEEE Conference on*, IEEE, London, UK, pp. 114-121.

Crosby, P., 2015. "Crowd-Sourced Stock Sentiment Using StockTwits.", Online Available at https://www.quantopian.com/posts/crowd-sourced-stock-sentiment-using-stocktwits.

De Choudhury, M., Sundaram, H., John, A., and Seligmann, D. D. 2008. "Can blog communication dynamics be correlated with stock market activity?," in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, ACM, Pittsburgh, PA, USA, pp. 55-60.

Fama, E.F. 1970. "Efficient capital markets: a review of theory and empirical work," *Journal of Finance* (*25)*, pp. 383–417.

Janetzko, D. 2014. "Predictive modeling in turbulent times–What Twitter reveals about the EUR/USD exchange rate," *Netnomics: Economic Research and Electronic Networking* (15:2), pp. 69-106.

Keynes, J. M. 1935. "The General theory of employment, interest and money," *Macmillan Cambridge University Press, for Royal Economic Society.*

Medhat, W., Hassan, A. and Korashy, H. 2014. "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, 5(4), pp.1093-1113.

Melville, P., Gryc, W. and Lawrence, R.D. 2009. "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,* ACM, pp. 1275-1284.

Nann, S., Krauss, J., and Schoder, D. 2013. "Predictive Analytics on Public Data – the Case of Stock Markets," in *Proceeding of the 21st European Conference on Information Systems (ECIS),* Utrecht, NL.

Oliveira, N., Cortez, P., and Areal, N. 2013. "On the predictability of stock market behavior using stocktwits sentiment and posting volume," in *Correia L., Reis L.P., Cascalho J. (eds) Progress in Artificial Intelligence. EPIA 2013. Lecture Notes in Computer Science, vol 8154.* Springer, Berlin, Heidelberg, pp. 355-365.

Oh, C., and Sheng, O. 2011. "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement," *Thirty Second International Conference on Information Systems ICIS*, Shanghai, pp 1-19.

Papaioannou, P., Russo, L., Papaioannou,G., and Siettos, C. I. 2013. "Can social microblogging be used to forecast intraday exchange rates," *Netnomics: Economic Research And Electronic Networking,* (14:12), pp. 47–68.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. 2015. "The effects of Twitter sentiment on stock price returns," *PLoS ONE 10(9): e0138441.*

Rao, T., and Srivastava, S. 2014. "Twitter sentiment analysis: How to hedge your bets in the stock markets," In Fazli Can, Tansel Özyer & Faruk Polat (Eds), *State of the Art Applications of Social Network Analysis* (227-247). Springer International Publishing.

Sabherwal, S., Sarkar, S., and Zhang, Y. 2008 "Online talk: does it matter?" *Managerial Finance* 34(6):423-436.

Sprenger, Timm O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. 2014. "Tweets and trades: The information content of stock microblogs," *European Financial Management* 20.5 (2014): 926-957.

Tumarkin, R., and Whitelaw, R. F. 2001. News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51.

US Securities and Exchange Commission 2013. "SEC Says Social Media OK for Company Announcements if Investors Are Alerted," Online available at *http://www.sec.gov/News/PressRelease/Detail/PressRelease/1365171513574.*

Veiga, Andre, Peiran Jiao, and Ansgar Walther. "Social Media, News Media and the Stock Market," *News Media and the Stock Market* (March 29, 2016) (2016).

Appendix

List of terms used in Naïve Bayes classifier for identifying communication about EUR/USD

EUR/USD, EURUSD, EUR\s*USD (regular expression), $EURUSD (Tweets only), $EUR/USD (Tweets only), Euro, US Dollar, ECB, FED, FOMC