Lund University School of Economics and Management
Master's Thesis in Finance (NEKN02)
Supervisors: Anders Vilhelmsson and Sara Moricz

Twitter Sentiment and Stock Returns

Investigating Twitter sentiment forecasting power

by

Fredrik Wolf fredrik.wolf@me.com
Oskar Bergdorf oskar.bergdorf@gmail.com

Abstract This study aims to investigate if the sentiment expressed on Twitter has an effect on individual stock returns. We use a uniquely large data-set consisting of 129 million tweets related to 31 companies over a 15-month period. The sentiment is derived on a daily basis with Loughran and McDonald's established finance-focused method, as well as with Vader, a sentiment analysis tool developed specially for social media. Using a panel data regression, we empirically test Twitter sentiment effectual forecasting power on individual stock returns. Our main findings are: 1) the sentiment derived from Twitter can help predict individual stock returns up to two days ahead 2) the individual stock returns are more sensitive to 'cashtag'-tweets than general-company tweets 3) Twitter data can be used to explain the stock return volatility 4) Loughran and McDonald Lexicon-based method outperforms Vader. The results indicate that Twitter data is a suitable data source to understand and forecast stock market movements.

Keywords: sentiment analysis, Twitter, stock return, volatility, prediction, forecasting

Acknowledgements

We express our gratitude to Sara and Anders for their valuable advice and encouragement during this project. We also want to thank our friend Fritjof for lending us the computational power needed to complete this project.

Contents

List of Figures

List of Tables

1 Introduction

The ability to predict asset prices would be highly valuable for investors and other parties. Consequently, the field of stock market prediction has received a great deal of attention from both academia and businesses. The early research on the prediction of the stock market was based on random walk theory and the widely known Efficient Market Hypothesis (EMH) (Cootner, 1964; Fama, 1965). According to the EMH, all the available information is already reflected in the asset prices and should only react when new relevant information about the asset emerges. The stock market prices will therefore follow a random walk pattern and cannot be accurately predicted, since news is unpredictable (Qian and Rasheed, 2007; Fama, 1991).

However, several studies have questioned the basic assumptions in EMH and found that the stock market prices can in fact to some degree be predicted by showing that the prices do not follow a random walk (Butler and Malaikah, 1992; Qian and Rasheed, 2007). Recent research also suggests that even though news may be unpredictable, early indicators that may predict changes and outcome in various socio-economic events can be extracted from online social media. Tumasjan et al. (2010) show a clear relationship between the number of times a particular party was mentioned on Twitter and the outcome of the German federal election. Asur and Huberman (2010) find that movie box offices can be predicted by extracting the public sentiment from the microblogging platform Twitter.

Behavioural finance research shows that our emotion and mood plays a large and important role in our financial decision-making (Nofsinger, 2005). Therefore, it is reasonable to assume that the public sentiment, mood and opinion may affect the stock market prices and movements. Recent studies support this, Edmans et al. (2007) show that a loss in a soccer game had a significant market effect. Hirshleifer and Shumway (2003) assess the public mood from the weather condition and find that sunshine has a significant correlation with stock returns.

The degree to which these methods accurately indicates the public mood can be questioned and with the recent increase of opinionated data online on social media there has been a significant progress of the techniques for gathering the public sentiment (Pak and Paroubek, 2010). The social media platforms are powerful tools for the user to share their opinions, emotions and attitudes which may provide a better representation of the public sentiment and mood. The microblogging platform Twitter has received much research attention after a study published by Bollen et al. (2011). The authors find that certain types of moods and sentiment extracted from Twitter have a significant predictive power of the Dow Jones Industrial Average

(DJIA) returns and an 87.6 % accuracy in predicting the direction for the returns (Bollen et al., 2011).

While several studies find a correlation between the sentiment on Twitter and market indices, there is little in the literature about individual stock returns. Two studies that test this on company level are Smailović et al. (2013) and Sprenger et al. (2014). They use a relatively small set of 0.15 and 0.25 million tweets to show that there is a significant predicting power for individual stock returns.

Prior research also shows that Twitter sentiment and data may affect the stock market volatility. In a study by Zhang et al. (2011) the authors find a positive relationship between emotional tweets and periods of uncertainty. Sprenger et al. (2014) expands on this research and show that disagreement in the public mood extracted from Twitter influences the trading volume. Since trading volume and volatility tend to move together, it is possible that user postings on Twitter might help forecast the market volatility (Antweiler and Frank, 2004).

This paper aims to add to the existing research on the predictability of stock returns and address the limitations of prior related work by making the following contributions. Firstly, only a few recent studies have taken the first step in exploring Twitter sentiment information with respect to individual stock returns rather than broader stock market indices. With our uniquely large data-set of 129 million tweets, we conduct the most comprehensive study to date on predicting individual stock returns using sentiment analysis.

Secondly, our paper is to the best of our knowledge the first that use two different data-sets of tweets, cashtags (tweets with more focus on the company stock) and general-company (tweets relating to the general company, its products or services). This allows us to compare the predictive validity of peoples general opinion about a company with the sentiment surrounding the company stock.

Furthermore, we test a new sentiment analysis tool named VADER that is constructed specially for social media sentiment analysis. VADER has to the best of our knowledge not been used for financial forecasting. We also compare VADER with the popular and frequently used sentiment analysis method LM Loughran and McDonald (2016) that include many finance specific terms. This allows us to determine the essential features of the sentiment analysis tool when forecasting stock returns using social media.

Lastly, we investigate if the number of tweets and the sentiment variance can help predict the volatility of the stocks. This extends on the previous research that suggests that Twitter data can help explain certain stock market features.

The rest of the paper is outlined as follows. Chapter 2 presents the main empirical findings in previous related research in the field of sentiment analysis and its connection to asset prices. Chapter 3 discuss the data collection and the data preparation. Chapter 4 describes the main methodology used in this paper. Chapter 5 presents the result from our tests. Chapter 6 concludes and discuss possible improvements and future research on the topic.

2 Theory

2.1 Existing literature and theoretical framework

A central concept when investigating the possibility to predict stock market returns is the efficient market hypothesis (EMH). The concept was first developed by Fama (1965) and states that all the relevant available information is already reflected in asset prices. The implication is that it is impossible for an investor to consistently beat the market on a risk-adjusted basis. The only way for an investor to gain a higher return than the market would be to invest in riskier assets (Fama, 1965). A great deal of research has been dedicated to testing EMH. The result is inconclusive as some of the studies supports it while others reject the theory (Butler and Malaikah, 1992; Lo and MacKinlay, 2002; Qian and Rasheed, 2007).

The noise trading theory suggests that there are prolonged market inefficiencies in asset prices. It states that the mispricing of an asset only rebounds to the fundamental value after a while and that the reason for this is that investors can lose money in the short term if they correct the market straight away (De Long et al., 1990). Hong and Stein (1999) explains that this 'noise trader risk' is especially noticeable in smaller firms since the incorporation of relevant information is slower than for larger firms. Although the theory explains why there is mispricing of assets over certain time periods, it does not mention how the mispricing happens in the first place.

In a study by Barberis et al. (1998), the authors try to explain how people form their expectations and beliefs and how it may relate and lead to the mispricing of assets. They use two known behavioural biases. The first is representation bias, which refers to how peoples' prior beliefs about an asset largely influence the formation of new beliefs and opinions when new information emerges. This may lead to investors neglecting the actual probabilities of certain scenarios. The second is conservatism bias, which suggests that the public's expectation and beliefs are slow to react to new information. This implies that even when investors update their investment strategy in the correct way, they do it in a smaller magnitude (underreaction). (Barberis et al., 1998)

Griffin and Tversky (1992) propose a theory that combines these two different biases and suggests that peoples decision-making is based on weight and strength. Weight refers to the statistical properties of the single event, while strength is the salience and emotional importance of the news or event. According to the authors, people tend to put too little importance on the 'weight' of the news/event and substantially more on 'strength'. This causes underreaction when fundamentally important news (high strength) are not that emotionally important.

2.1.1 Sentiment from different mediums and asset prices

There has been an increase in research on the relationship between asset prices and sentiment in recent years since the amount of easy access to opinionated data has increased substantially (Liu, 2012). In this section, we will provide the reader with an overview of the previously used mediums for sentiment and the results of the papers. We will end the section with some papers that used social media as their source of sentiment and focus especially on the studies with Twitter data. Kearney and Liu (2014) provides an extensive summary and comparison of the sentiment sources that are used in the previous literature. They find the most used sentiment sources in the field are from media articles, public corporate disclosures and Internet messages.

The media articles refer to news articles with relevant information that focuses more on reporting information than expressing opinions. Kearney and Liu (2014) therefore argues that the information provided in this medium is related to past and current events and not as much to the future prospects of the assets. In a study conducted by Tetlock (2007), the media sentiment source is used to show that pessimism in the media can predict negative index market returns. The authors conclude that the results are in line with the noise trading theory as the effect revert to the fundamental value over a few days and that it takes longer for a smaller firm.

The authors argue that corporate disclosures come with fundamental information embedded with sentiment. As previously discussed, the presence of fundamental information exposes the medium to the possibility of underreaction or overreaction. The low frequency of reporting corporate disclosure is a disadvantage which has led to that most of the research conducted with this medium are event studies that focused on different types of corporate reports. The main findings of these studies are that changes in mood from the previous corporate disclosure have a significant effect on asset prices even after controlling for possible surprises in the fundamental information. (Kearney and Liu, 2014)

In the internet medium, we have message boards, blogs and microblogs that all have very different characteristics than the two mediums mentioned above. In this medium, there is a substantial amount of noise in the new relevant information. For this reason, Kearney and Liu (2014) argues that it is a very interesting source to research market inefficiencies and confirm some of the behavioural financial theories. In addition to this Barberis et al. (1998) argues that the combination of the users' ease of spreading their opinion (high salience) with the lack of relevant fundamental information may lead to investors overreacting more than they underreact. Another study suggests using social media to extract sentiment as it is a better predictor than conventional media of stock returns (Yu et al., 2013). To summarize, the internet medium and social media in particular therefore seem to be the most relevant medium of the three for behavioural sentimental analysis and this is why we have chosen to use social media as our sentiment source.

2.1.2 Sentiment analysis on stock market indices

Antweiler and Frank (2004) study message boards postings on Yahoo!Finance and Raging Bull in order to predict market returns. They find evidence that positive postings can predict negative returns for the following day. The authors also use a more uncommon intraday data with 15 minutes interval to show that there is statistically significant predicting power even with a shorter interval.

Bollen et al. (2011) is one of the first well-cited studies that researched the connection between sentiment on Twitter and the stock market returns. In their study, they collect 9.9 million tweets to derive the public mood during a ten month period in 2008. They categorize the tweets using a dictionary-based approach into different mood states: anxiety, confidence, calmness, energy, happiness and kindness. By running time-series regression, the authors of the paper show that some mood dimensions have predicting power over the DJIA returns. They find that the dimension 'calm' and that changes in the public mood are significant predictors for the DJIA return up to 5 days in advance. 'Calm' has an accuracy of 88% when predicting three days ahead return of the DJIA index. The authors also divide their tweets into positive and negative and apply a Granger Causality test, their result once again shows that sentiment from tweets can be used to predict DJIA index returns. A similar study done by Mittal and Goel (2012) use Twitter data and previous days' DJIA values to forecast future stock movements. The authors use a data-set of 475 million tweets over a time frame of 7 months in 2009. Similar to Bollen et al. (2011) they categorize four mood classes for their sentiment analysis and then use both a dictionary based and a machine learning technique to test the relationship between the DJIA index and the public mood captured on Twitter. The paper confirms that 'calm can be used to predict DJIA returns but also finds that 'happy' have significant predicting power.

Zhang et al. (2011) use a different approach by testing if the emotional attitudes: anxiety, hope, happiness, fear and nervousness can predict market returns. They use a six-month sample and 5.5 million tweets in their study. The authors find that emotional tweets, unlike neutral tweets, are in general negatively related to the on-day-ahead values of the stock indices DJIA, NASDAQ and S&P 500. They also show that the total number of followers and retweets have a potential predicting power of indices values but is lower than the number of emotional tweets. In addition to this, the authors find that emotional tweets have a positive relationship to VIX (which is a stock market expected volatility) which implies that users on Twitter tend to use more emotional words, both positive and negative at time periods of increased uncertainty.

2.1.3 Twitter sentiment on individual stock return

A great deal of research in this field has been conducted on Twitters relationship to market indices such as DJIA and S&P 500. However, there is much less research done on individual company returns. In the section below we present some papers that research this area.

In a study by Ruiz et al. (2012), the authors look at 150 companies from the S&P 500 index. From the collected tweets, the authors then extract certain features which they want to test for correlation with the stock market. They show that there is a small correlation between sentiment and the returns of the individual firms. Despite this small correlation, they are able to propose a profitable trading strategy from their results. In addition, the author finds that the sentiment correlates with the trading volume.

Smailović et al. (2013) research how sentiment can be used to predict individual stock returns. They use a sample of 8 companies and 0.15 million tweets and a machine learning approach to extract the sentiment. By using a linear regression, they then find that changes in positive sentiment have significant predicting power for stock returns. This is especially the case when there are high variations in stock prices or a significant fall in the returns. Tweets that are classified as neutral can provide additional information in some cases when modelling stock returns. Sprenger et al. (2014) have a similar approach as Smailović et al. (2013) except that they use panel data. By using a sample of 0.25 million tweets, they find that there is a significant relationship between bullishness in twitter data and company returns.

2.1.4 Volatility

Several of the studies mentioned above also find interesting result regarding social-media and its relationship to market features such as trading volume and volatility of the stock market.

Antweiler and Frank (2004) study how postings on Yahoo!Finance and Raging Bull may relate to the stock market features. The authors find a strong positive correlation between the amounts of postings and trading volume. They also show that the number or message postings and the volatility are positively correlated. Furthermore, they find that agreement in the message postings has a positive correlation with the volatility. This is interesting since some financial theories and empirical evidence imply that disagreement between market participants should induce trading and thus increase volatility (Harris and Raviv, 1993). However, the result is in line with what Das et al. (2005) and Danthine and Moresi (1993) suggest. The authors of these papers argue that agreement in the market may lead to less information being released due to the lack of extensive debates. Less available information decreases rational agents chances to counteract noise traders and their actions, thus increasing the volatility.

Sprenger et al. (2014) also extends on Antweiler and Frank (2004) study by analysing the microblog's association with the stock market features. Similar to Antweiler and Frank (2004), the authors find a strong positive correlation between the number of posts on Twitter and volatility. Interestingly the authors also find that disagreement in the public mood increases trading volumes. This is in line with the financial theory that suggests disagreement among market participants causes the trading volume to rise since the assets are differently valued by the market Harris and Raviv (1993) Karpoff (1986) .

Furthermore, in a study by Zhang et al. (2011), the authors test if the number of emotional tweets correlates with VIX (which is an index for the expected stock market volatility). They find that during times of higher volatility and uncertainty, people tend to express themselves with more emotional words. Another study by Ruiz et al. (2012) find that the sentiment collected from tweets correlates with the trading volume.

Table 2.1: Litarature review

In this table we present an overview of the most relevant literature, that covers the relationship between asset returns and Twitter data. The time frame column denotes the time period used for study as well as it's length. The data column contains data frequency, size of data-set, as well as financial data. In the methodology column, the reader can find which sentiment analysis method the authors use, as well as how the relationship is modelled. We provide key takeaways from each study in the results column.

#	Authors	Time Frame	Data	Methodology	Results
1	Bollen et al. (2011)	10 months, 2008	Daily: 9.9 m tweets; DJIA	Dictionary-based; Linear regression + SOFNN*	The mood dimension 'calm' have predicting power of DJIA returns and that change in the mood are significant predictors for the DJIA return up to 1 to 5 days inadvance.
2	Mao et al. (2011)	15 months, 2010-2011	Daily and Weekly: Tweets, Google search, Surveys, News; DJIA, VIX, gold	No sentiment analysis; Linear regression	Show that the sentiment from all mediums were correlated with both daily returns and VIX. They also find that the Twitter bullishness and the amount of tweets have significant predicting power for returns on both 1 and 2 lags. The data from News have lower significance and survey data was not significant. On a weekly basis the Google search volumes have significant results.
3	Zhang et al. (2011)	5 months, 2009	Daily: 5.5 m tweets; DJIA, S&P 500, NASDAQ	No sentiment analysis; Correlation	The word with the most significant next day return were words such as worry, fear, anxious and hope. The authors found that a combination of negative words have a stronger predicting power than combination of positive words and volatility. There is a positive correlation between number of emotional words and the next day volatility but a negative correlation for the next day returns.
4	Mittal & Goel (2012)	7 months, 2009	Daily: 475 m Tweets; DJIA	Dictionary-based; Linear regression + SOFNN*	Twitter data can capture the public mood and find that 'happy' and 'calm' has a 3-4 lagged relationship with DJIA. They find that the SOFNN* outperforms the other methods.
5	Chen and Lazer (2013)	Not presented	Daily: Tweets; Market returns	Dictionary-based; Linear regression	Find that the sentiment on Twitter correlates with the market returns. They purposed several profitable trading strategies by incorporating the sentiment data.
6	Smailovic et al. (2013)	9 months, 2011	Daily: 0.15 m tweets; 8 companies	Machine learning; Linear regression	Changes in positive sentiment have a significant predicting power for stock returns. This is especially the case when there are high variations in stock prices or a significant fall in the returns. Tweets that are classified as neutral can provide additional information in some cases when modelling stock returns.
7	Yu et al. (2013)	3 months, 2011	Daily: 0.05 m messages from companies, forums, media and Twitter; 824 companies	Dictionary-based; Panel regression	Both conventional media- and social media sentiment show a significant effect on individual stock return, the effect is stronger for social media sentiment. There is a positive effect on risk for both blogs and Twitter sentiment. Find that sentiment on forums correlates negatively with return.
8	Sprenger et al. (2014)	6 months, 2010	Daily: 0.25 m tweets; Individual companies	Machine learning; Panel regression	Significant relationship between bullishness in Twitter data and stock returns. Variance in the public mood is associated with an increase in the trading volume.
9	Mao et al. (2015)	36 months, 2010-2012	Daily: 0.31 m messages from Twitter and Google search: DJIA, S&P 500, Russell 1000 and 2000	No sentiment analysis; Linear regression	Twitter outperformes Google search queries in estimating stock market sentiment. Bullishness on Twitter is able to predict index returns in Canada, UK, and US. Within a week the prices reverts to fundamentals.

*SOFNN is a self-organized fuzzy neural network, a subset of machine-learning algorithms.

2.2 Summary and research questions

To summarise, previous research finds a statistical association between sentiment on Twitter and the stock market. The methods used in the literature differ and there does not seem to be a preferred way to do the sentiment analysis, but the two most common is the machine learning and dictionary-based approach. Linear regression is then often used to find the relation with the stock market returns. Several studies find strong same day correlation and some studies even find significant predicting power. The most common finding in the literature is that sentiment has significant predicting power for the one-day-ahead return. Bollen et al. (2011) also show that the twitter sentiment has predicting power for certain indices up to five days ahead. Some of the previous research is in line with the noise theory as Mao et al. (2011) shows that the sentiment has significant predicting power, but also that returns reverse to the fundamental values within a week. The studies that focus on predicting individual stock find a similar result as the studies that research the market returns (Smailović et al., 2013). Several studies have shown that trading strategies that use Twitter sentiment lead to higher portfolio returns, which is a strong indicator that there is an economic significance in the results.

Prior related work has also shown that it might be possible to predict market features such as trading volumes and volatility using data from social media. Previous research find that there is a strong positive correlation between the amounts of user post and the trading volume in the following day Antweiler and Frank (2004); Sprenger et al. (2014). In addition to this Antweiler and Frank (2004) shows that agreement in the postings correlates positively with volatility.

The research between social media and stock market returns are rapidly growing. We aim to contribute to the literature by testing the previously found relationship between Twitter sentiment and index returns on a company level. We also test if recent advances in sentiment analysis can help improve the results. Based on previous research, the following hypotheses will be tested in this paper:

H1: *Twitter sentiment has predictive power on company returns.* Common hypothesis for market indices returns, but only a few well-cited studies have tested this on individual stock returns. These studies have used a very small sample of both tweets and companies. We, therefore, add to current research predicting individual stock using sentiment analysis with our data-set consistent of 129 million tweets and 31 companies from S&P 500. In addition, we want to contribute to the method of extracting sentiment since there seems to be no consensus in prior research on what is best. Hence, we provide the reader with a comparison of the two different sentiment analysis tools VADER and LM.

H2: *Returns are more sensitive to cashtags*. This hypothesis has to our knowledge not been tested in prior related studies. Mao et al. (2012) argues that general-company tweets include more information and thus might be better at predicting stock returns. However, Sprenger et al. (2014) oppose this and argues that there is too much unrelated noise in the general-company information set. We want to add to this discussion by investigating which of the two is actually the best for predicting stock returns. We believe that for the daily return prediction, there will be too much noise in the general-company set, for it to have any significant advantages.

H3: *Twitter sentiment and data can be used to explain volatility*. On this topic, there is very little previous research and we believe that this hypothesis can help increases the understanding of how the public sentiment influences the stock market and its ability to predict market movements. We test if the number of tweets, due to its previously found relationship with trading volume, can help capture and explain volatility. We also believe due to previous findings that the variance (disagreement) in the public sentiment can help explain volatility.

3 Data

3.1 Twitter data

We collect original public tweets from Twitter for the period 2017-12-31 to 2019-03-31. For each tweet we have its tweet identifier, date and time of creation, tweet text, number of likes, number of retweets and number of replies. The tweets were collected using search queries for the related companies. In line with previous research, we decide to collect a data-set of tweets by company 'cashtags' (Mao et al., 2012). Similar to hashtags, in 2012, Twitter introduced the use of a dollar sign ($) preceding a company's ticker symbol for users to denote they were talking about a specific stock (i.e. Microsoft's cashtag is $MSFT) (Daniel et al., 2017; Mao et al., 2012; Smailović et al., 2013). Building on this, we follow the recommendations of previous research to extend the data-set to include more search queries for a larger data-set (Daniel et al., 2017; Mao et al., 2012). For this second data-set we collect tweets using company names as search queries (a comprehensive list of search queries is presented are Tables A.1 and A.2, Appendix). It should be noted that both the public opinion of a stock, as well as the more general public opinion of a company and its offerings should have an impact on the company's returns (Pagolu et al., 2016). However, the sentiment derived from the general-company tweets may be a noisier signal, as it also contains more unrelated tweets.

We collect the tweets for the 30 largest companies on the S&P500 index as well as for Tesla, Inc. We include Tesla in the sample as we find it interesting to include a company where its management are famous for their use of Twitter. For the second data-set, we omit Apple due to the generic nature of its name. In addition, we opt to use Alphabet's old name for tweet collection, as the company is still mostly referred to as 'Google', in conjunction with Alphabet being a very generic term, too. While collecting the tweets we filter out any non-English tweets, as our sentiment analysis tools have mainly been trained and created for English texts. Moreover, for the first data-set, we drop any tweet containing more than one cashtag, so that ambivalent tweets discussing more than the intended stock will not be accounted for, as well as to remove potential spam (tweets with excessive use of cashtags). We end up with $\sim$3.5 million tweets for the first data-set and $\sim$125.7 million tweets for the second data-set, a substantially larger data-set than that of previous studies on company-level, as seen in the literature review in Table 2.1.

3.2 Financial data

We collect daily adjusted closing prices for the 31 companies in our sample for the period between 2018-01-01 and 2019-03-31 from Yahoo!Finance (2019). In addition, we collect the daily adjusted closing value for Standard&Poors 500 index (S&P500) for the same time period. We then calculate the daily log-returns in accordance with previous studies (Mao et al., 2011):

$$r_t = \ln\left(\frac{S_t}{S_{t-1}}\right) \tag{3.1}$$

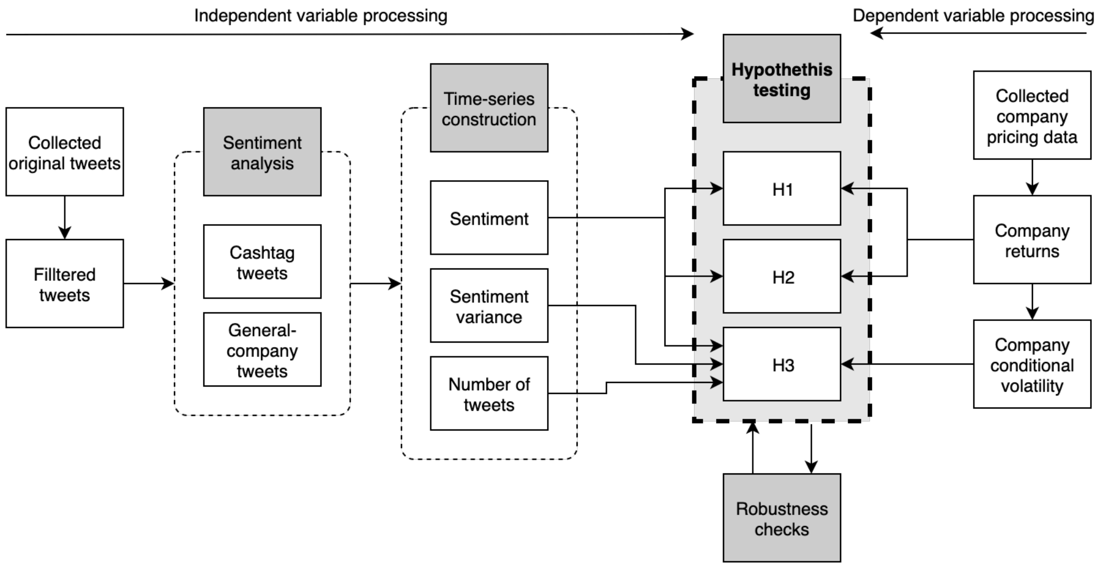
where r_t is the daily log-return on day t and S_t is the adjusted closing price on day t.

4 Methods

In this section, we discuss further data-processing, sentiment analysis, sentiment time-series construction and empirical testing methods. An overview of the steps is shown in Figure 4.1.

Figure 4.1: Steps of collecting and processing data

The figure depicts the processing of Twitter and financial data before hypotheses testing. In the previous section, we describe the initial data collection process, including Collected original tweets, Filtered tweets, Collected company pricing data as well as Company returns. The next step for the independent variable processing is sentiment analysis, which is further explained in Section 4.1. We do this for both data-sets before continuing to the time-series construction, described in Section 4.2. The conditional volatility processing is explained in Section 4.3.2. Finally, the processed variables are used in the regressions, which is further explained in Section 4.3.



4.1 Sentiment analysis on Twitter data

Sentiment analysis (also opinion-mining or emotion AI) is an area in the field of natural language processing, that uses computational linguistics and textual analysis to systematically extract, quantify and analyze affective states and subjective information from a text (Pang et al., 2008; Liu, 2012; Wilson et al., 2005). There are two main approaches for textual sentiment classification: machine-learning and dictionary-based analysis (Kearney and Liu, 2014). The machine-learning technique classifies sentiment based on previous dynamics and patterns from a sample data-set. The dictionary-based approach classifies sentiment based on a pre-defined dictionary. Thus, the results rely heavily on the quality of the dictionary as well as how the words are weighted. We have decided to use the dictionary-based approach for two reasons. Firstly, previous research suggests that there are no significant advantages using a machine learning algorithm over the simpler dictionary-based approach for data classification, especially for social media sentiment classification (Hutto and Gilbert, 2014). Secondly, there is no pre-defined data-set of tweet data we can use to train the machine learning algorithm on. There is no consensus regarding which dictionary-based sentiment analysis technique performs best (Hutto and Gilbert, 2014; Ribeiro et al., 2016; Araujo et al., 2016); it depends very much on how the dictionary was created, for what medium and purpose. As we have a financial focus for our sentiment analysis, we decide to use two methods for deriving sentiment of tweets: a finance-focused lexicon by Loughran and McDonald (2016) and VADER (Valence Aware Dictionary and sEntiment Reasoner) a sentiment analysis tool developed specifically for shorter social media texts (Hutto and Gilbert, 2014).

Researchers commonly use external word lists such as Harvard's General Inquiry Dictionaries (2019), which has the benefit of its content lying outside the control of the researcher. Loughran and McDonald (2016), however, find 73.8% of negatively classified words are not typically negative in a finance context. Using a large sample of 10-Ks fillings, Loughran and McDonald (2016) extend the Harvard/GI negative words list with finance-specific terms to increase its accuracy. Their dictionary (LM) is now one of the most used methods in the financial literature (Cortis et al., 2017; Mao et al., 2011), and has subsequently been shown to outperform various other commonly used methods of classifying textual data (Li et al., 2014). A disadvantage of using LM on social media posts is that it was originally created for use on longer texts, hence it may prove sub-optimal for analyzing tweets. Another shortcoming is the fact that LM does not account for negation (i.e. 'is not good' will be counted as positive, rather than negative). To counteract this problem, and in accordance with Loughran and McDonald's (2016) suggestions for improvement, we implement a basic negation for positive words, using 'negation words' provided in their master dictionary updated for 2018 (Loughran and McDonald, 2019). Our implementation reclassifies a positive word into a negative word if it is preceded within three words by a negation word. Finally, we used the established method of calculating sentiment polarity (Twedt and Rees, 2012; Kearney and Liu, 2014):

$$LM_i = \frac{N_{i,pos} - N_{i,neg}}{N_{i,pos} + N_{i,neg}} \tag{4.1}$$

where LM_i is the LM sentiment polarity score for tweet i, $N_{i,pos}$ is the number of positive words and $N_{i,neg}$ is the number of negative words in tweet i. As it is a polarity, we get a score between -1 and 1, where -1 should interpret as extremely negative and 1 as extremely positive, leaving a neutral tweet around 0. However, due to the limited amount of words in each tweet, we find a lot of tweets end up on the extreme sides of the scale.

VADER, on the other hand, is slightly more sophisticated. It has a few advantages over other conventional models. Most notably it handles:

(i) Pre-defined treatment of negation

(ii) Punctuation, i.e. three exclamation marks increase the strength of text more than just one exclamation mark

(iii) Capitalization, writing in capital letters changes the strength

(iv) Constructive conjunctions, such as 'but'

(v) Strengthening adverbs such as 'awfully good'

(vi) Emojis and emoticons

VADER has since its inception outperformed many well known dictionary-based approaches, as well as machine-learning techniques (Hutto and Gilbert, 2014; Ribeiro et al., 2016; Araujo et al., 2016). However, to the best of our knowledge, VADER has not been used in any well-cited paper for financial modelling.

VADER provides positive and negative valence scores as well as a normalized, weighted composite score for each tweet. We use the latter, as a uni-dimensional measure of sentiment is the most useful metric for our research. The composite score is, similar to LM, normalised to be between -1 (extremely negative) and 1 (extremely positive). Furthermore, we follow Hutto and Gilbert's (2014) recommendation of classifying tweets with a score ≤ -0.05 as negative, $-0.05 <$ score < 0.05 as neutral and $0.05 \leq$ score as positive[1]. Table 4.1 highlights some of the differences in scoring between the two, we can for example see in tweet 4 that LM is able to score 'overvalued' negatively, while VADER fails to do so.

4.2 Sentiment time-series

After we assign each tweet a sentiment score $(s_{t,j})^2$, we aggregate the sentiment scores to form a time-series of daily sentiment $(SENT_t)$. In order to form the time-series, we must first define time-thresholds (T and T-1) as well as the aggregation method. The literature is not consistent when it comes to the definition of time-thresholds. We define it as the stock-exchange closing hours, meaning that all tweets

[1] There is no consensus in the literature regarding thresholds for negative and positive score. We test using the more conservative thresholds of $+/-0.25$ and $+/-0.5$. However, we do not find any improvement to our results. Thus, we fall back on Hutto and Gilbert's (2014) recommendation. We use the same threshold for the LM method to keep consistency within our paper.

[2] As discussed earlier, we actually have two sentiment scores for each tweet, LM and VADER, however, we omit this notation for brevity going forward.

Table 4.1: Example of tweets and their sentiment scores

This table presents some randomly chosen tweets and their respective sentiment scores, as scored by VADER and LM. Tweets 1 to 5 are from the cashtags data-set and tweets 6 to 10 are from the general-company data-set. Both sentiment scores range from -1 to 1, where -1 is extremely negative and 1 extremely positive. Scores around 0 are considered neutral in sentiment.

		Sentiment score	
#	Tweet	VADER	LM
1	$msft looking very nice! wish i held a tad longer	0.7479	0.0000
2	yaya $msft reversing finally	0.0000	0.0000
3	$MSFT wow	0.5859	0.0000
4	$MSFT one of the few that has not taken an hit since many years...grossly overvalued [image redacted]	0.0000	-1.0000
5	$MSFT all time highs and we in a damn bear market? You poor fkn shorts good lawd...	-0.4404	0.0000
6	Bill Gates meeting set with President Trump - CNET [link redacted] #microsoft	0.0000	0.0000
7	#E2 is over. Everything we have learnt however will continue through us and to our students. TY to Microsoft for the amazing experience, and big TY to all the people that worked tideously to make the event a success. #RoadBackFromE2	0.8964	1.0000
8	2- After about 15-min of chat, CSR told me they'd escalate becuase they can't do anymore than they did, which is having me re-install Skype!!! After investigating on my own, I was able to find the security page buried very deep in Microsoft website (Thanks UX team)	0.5067	-0.3333
9	Teacher in Ghana who used blackboard to explain computers gets some Microsoft love - [link redacted] #TechNews [image redacted]	0.6369	0.0000
10	@Microsoft STOP UPDATING OUR SERVERS!!! YOU ARE SERIOUSLY SCREWING UP!!! YOU FAILED AT YOUR JOB AND AT LIFE!! #nomoreautoupdates	-0.9217	-1.0000

created after the stock-exchange closes will be aggregated into the next trading day. We use the exact trading time, taking public holidays and half-days into account. The key reason for this choice is that, similar to conventional mediums, we expect new information in tweets created outside of trading time to be reflected in the stock price in the following trading day.

In regards to the aggregation method, we follow previous literature. Antweiler and Frank (2004) discuss this area in depth and suggest three different methods. We chose to use two of these[3], which we find are the most appropriate for our research as well as present one of our own. These are presented in Equations[4] 4.2, 4.3 and 4.4. The first aggregation formula does not account for the number of tweets in a given day (count-neutral), while the second aggregation formula does. In it, the sentiment score is amplified by the number of tweets in that day (count-dependent). Furthermore, while Antweiler and Frank (2004) find similar results between the models, they also find the count-dependent aggregation method (Equation 4.3) slightly outperform the count-neutral one (Equation 4.2). This is consistent with other research, which find the number of messages does indeed have an effect on the results

[3]Note that Antweiler and Frank's (2004) research is conducted on message boards, using buy, hold and sell sentiment, rather than positive, neutral and negative as in our research. The notation presented in this paper reflects our research, thus it differs from that of Antweiler and Frank (2004).

[4]We do this for all firms, however, the firm index is omitted for notational convenience going forward where not needed.

(Mao et al., 2011)[5]. This is also the most commonly used method in the literature (Mao et al., 2015; Sprenger et al., 2014). The third aggregation formula is not commonly used in the literature. However, Smailović et al. (2013) discuss the potential of neutral tweets providing additional information when forecasting stock returns, which the first two aggregation methods completely discounts. In addition, VADER provides a more sophisticated, continuous sentiment score (as seen in Table 4.1), than that of previous research, we expect to lose efficiency by classifying a tweet with 0.5 as equally positive as one with a score of 1. By using the mean sentiment of each day, we capture these more intricate differences in sentiment provided by VADER as well as the information neutral tweets may contain.

Count-neutral aggregation formula:

$$SENT_t^A \equiv \frac{M_{t,pos} - M_{t,neg}}{M_{t,pos} + M_{t,neg}} \tag{4.2}$$

Count-dependent aggregation formula:

$$SENT_t^B \equiv \ln\left(\frac{1 + M_{t,pos}}{1 + M_{t,neg}}\right) \tag{4.3}$$

Mean sentiment aggregation formula:

$$SENT_t^C \equiv \frac{1}{M_t}\sum_{j=1}^{M_t} s_{t,j} \tag{4.4}$$

$M_{t,pos}$ is simply the number of positive tweets as defined by the threshold $0.05 \leq s_t$ and $M_{t,neg}$ is similarly the number of negative tweets defined by $s_t < -0.05$. M_t is the total number of tweets during day t. While it is possible to weigh each tweet by its retweets, replies or likes, as it might add additional information, we decide against it for two reasons. Firstly, these metrics are inherently lagged themselves, as their timing always follow the original tweets' creation. We are interested in the real-life application of our results. Thus it would be counterproductive to use such a weighing system. Secondly, Sprenger et al. (2014) find that there is no correlation between the quality of the information in a tweet and its number of retweets.

It shall be noted that previous research (Tetlock, 2007) find a negative sentiment to be a stronger predictor of negative returns than a positive sentiment is of positive returns. The time-series we construct do not take this 'leverage-effect' into account but instead focuses on the change in sentiment only. We realize the effect this choice

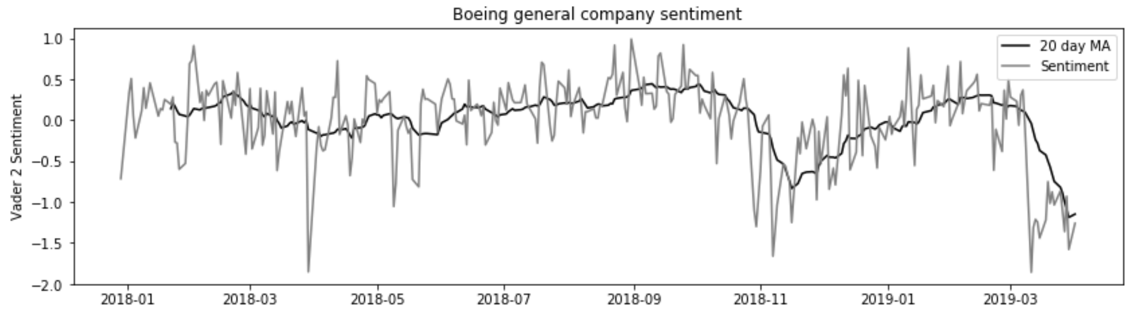
[5] After some investigation however, we find that the difference between the equations is highly dependent on the number of tweets. For example, the correlation between the two was 0.9040 for UnitedHealth's cashtags with 7,255 tweets (after filtering, using VADER) and 0.9997 for Tesla's cashtags with 362,495 tweets.

may have on our results, however, in order to keep comparability with other studies we still use the previously mentioned aggregation methods. A graphical presentation of an example sentiment time-series is found in Figure 4.2.

In addition to the sentiment time-series, we also form time-series of sentiment variance (as a measure for disagreement), the amount of negative, positive and total tweets for each day[6] to use in testing our H3.

Figure 4.2: Example sentiment time-series

This figure graphically presents the general-company sentiment for Boeing as scored by count-dependent VADER*. There is a very noticable decline in sentiment around the Lion Air Flight 610 crash on 29 October 2018, as well as around the Ethiopian Airlines Flight 302 crash on 10 March 2019. We present the 20-day sentiment moving average in addition to the raw sentiment for the reader's convenience, although we do not use the moving average in any of our methods.



*Note that the sentiment is demeaned.

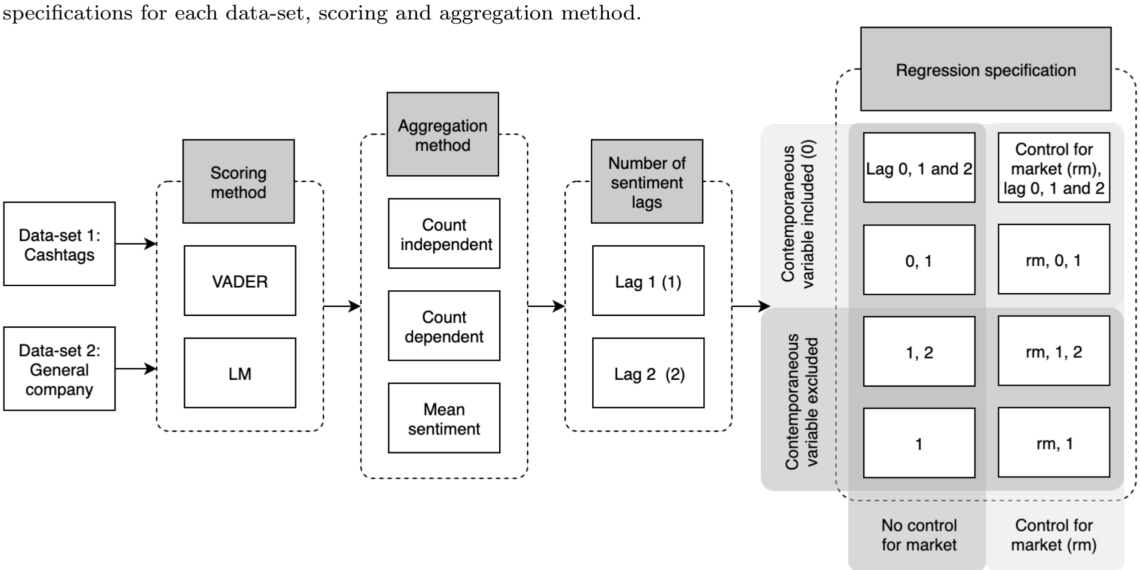
[6]We do this for both VADER and LM, so we have two time-series for the number of positive and negative tweets for each company.

4.3 Empirical methods

4.3.1 Return predictability

Figure 4.3: Steps of parameter estimation

The figure depicts the testing methodology, with the different regression specifications tested. We use eight different specifications. One and two lags, with and without the contemporaneous variable, as well as controlling for market return and not. The number in parenthesis depicts our naming-scheme for the regressions, which we use in our presentation in the results. We use Akaike information criterion to determine that two sentiment lags are optimal. 0 means contemporaneous variable included in regression, 1 and 2 means lag one and two of sentiment. rm means control for market returns. All regressions include the first order autocorrelation of return. We estimate these eight specifications for each data-set, scoring and aggregation method.



After we form the sentiment time-series for each company, it is time to test the relation between sentiment and company stock returns. There is no consensus in the literature as to which testing methodology is the best. Kearney and Liu (2014), however, provides a discussion regarding the different methods researchers use. The most commonly used method is ordinary linear regression (auto-regressive distributed lag model - ARDL). Some studies use VAR models to capture interdependencies and evolution between performance, controls and sentiment while, albeit less common, some use panel regression models to capture firm heterogeneity. Chen and Lazer (2013) discuss benefits of using simple linear regression, pointing out its speed when handling huge amounts of data, for trading strategies in real time (as is the case with twitter data), compared to more complicated models. In addition, a regressor provides valuable information regarding the level of change, as opposed to a simple classifier (i.e. logit or probit) which only provides the direction of change. Moreover, as Brown and Cliff (2004) find evidence of a two-way directional relationship between sentiment[7] and stock returns, as well other previous studies finding predictive power of Twitter sentiment on market-level returns (Bollen et al., 2011; Mao et al.,

[7] Albeit not using sentiment derived from Twitter.

2015), in order to focus on the directional relationship in line with our hypotheses (that is, the predictability of twitter sentiment over firm-level returns), we use a panel ARDL model with fixed entity effects in line with previous research on firm-level (Demers et al., 2008; Antweiler and Frank, 2004; Sprenger et al., 2014). We present its general form in Equation 4.5.

$$r_{t,i} = \alpha_i + \beta_i^r r_{t-1,i} + \sum_{j=0}^{n} \beta_{j,i}^{SENT} SENT_{t-j,i} + \sum_{k=0}^{1} \beta_k^{rm} rm_{t-k} \tag{4.5}$$

where r_t, rm_t and $SENT_t$ are the return, market return and sentiment on day t, respectively. In the panel model, we regress the return only on the sentiment time-series built using keywords for the related firm. In the literature, this is the most common method to predict returns, not only for individual stocks but for indices as well (such as the S&P500 or DJIA) (Sprenger et al., 2014; Ruiz et al., 2012; Smailović et al., 2013). Thus, we want to see if we find similar results as previous research with a new methodology and a newer and more comprehensive data-set.

There is no clear consensus on whether or not to include the contemporaneous variable. While Antweiler and Frank (2004) find significant intraday predictability, Brown and Cliff (2004) suggest it is a two-way relationship in the daily horizon. As the literature is ambivalent as to include it or not, we do not want to limit our scope in this regard. Thus, we estimate separate regressions with it included, and omitted. Regarding control variables, we include the first order autocorrelation of returns as well as the contemporaneous and one lag market return[8], which is common practice in the literature. We do, however, also estimate purely predictive models, in line with H1, without any contemporaneous explanatory variables. Other than that, the literature is very inconsistent as to which controls to include. Daily return predictability is highly complex, so this is to be expected. Furthermore, when significant explanatory variables are found, they become the basis for trading, which in turn causes the opportunities to be traded away.

4.3.2 Volatility predictability

Much less has been done in the volatility space in regards to Twitter sentiment, and a few different methods have been suggested. Moving forward, we must first define which metric to use for volatility, which sentiment features to test, as well as what estimation method. Most well-cited papers that model volatility do it on the market-level; thus, they have access to a wider range of volatility-metrics, such as the VIX-index (Zhang et al., 2011). On the company-level, some studies

[8]As discussed earlier, we use the S&P500 index as a proxy for the market.

use the implied volatility derived from option-pricing Oliveira et al. (2013), while some calculate realized volatility based on shorter horizon returns, or a GARCH-based approach for conditional volatility (Antweiler and Frank, 2004). Due to its popularity for modelling stochastic volatility, we use a GARCH-based approach. For each company is the sample, we estimate daily conditional volatility with a constant mean GARCH(1,1)-model. With the estimated conditional volatility, we then set up a GARCH-like panel model, with the Twitter sentiment-features of interest. We present its general form in Equation 4.6.

$$\sigma^2_{t,i} = \omega_i + \alpha_i \eta^2_{t-1,i} + \beta_i \sigma^2_{t-1,i} + \sum_{c=1}^{C} \gamma^c_i X^c_{t-1,i} \tag{4.6}$$

where σ^2_t, η_t and X^c_t are the estimated conditional volatility, residuals from GARCH-estimation and vector of Twitter sentiment-features c on day t, respectively. When it comes to the sentiment-features, a wide range have been suggested, but a few stand out. Disagreement among investors has long been considered a driving force behind trading motivation (Karpoff, 1986; Harris and Raviv, 1993), with some going as far as proposing theorems stating no trade would occur if a perfect agreement exists (Milgrom and Stokey, 1982). Some recent studies find evidence of investor disagreement being able to help predict volatility on market-level (Antweiler and Frank, 2004; Das and Chen, 2007; Chen and Lazer, 2013). We use the daily sentiment variance as a measure for investor disagreement, as more disagreement should, in turn, lead to higher variance in the sentiment. The raw sentiment is probably the most used variable in linking sentiment to volatility. Rao and Srivastava (2012) find it has predictive power, Das and Chen (2007) link sentiment to trading volume while Tetlock (2007) find predictability in the level of sentiment (positive or negative). Thus, we use the squared sentiment[9] in our volatility model. Finally, as Oliveira et al. (2013) and Zhang et al. (2011) find a significant relationship between the number of postings and volatility we test the total number as well as the amount of positive and negative tweets[10].

[9]Note that we de-mean the sentiment before squaring, as the sentiment is not zero-mean.

[10]We log-transform these as we are interested in the relative change of the number of tweets have on the volatility. Also, as we aggregate all tweets from closing time on Fridays to Monday, each Monday will have three days worth of tweets in its tweet-count. To counteract this, we divide the number of tweets with the number of days they were collected on, meaning that regular Mondays will actually have the average number of tweets per day over Saturday through Monday. We also do this for public holidays.

4.4 Diagnostical checks

In regards to the robustness of our results, we make sure the assumptions underlying our models are satisfied. Before running the regressions we test all variables for multicollinearity and stationarity. In addition, before running the panel regressions, we first run each specification for each company as an ordinary linear regression and test for autocorrelation with the Durbin-Watson test statistic. We opt for this method as the panel-regression implementation we are using in python does not support the Durbin-Watson test statistic.

In regards to multicollinearity, we use the rule of thumb of 80% and do not find any significant issues in our specifications. Moreover, we do not find any evidence of non-stationarity in the variables. For the regressions testing sentiment effect on returns, the Durbin-Watson statistic shows no evidence of autocorrelation (test statistic is *very* close to 2 for all companies and specifications). For the return panels, as we are working with financial data, we follow the consensus in the literature and calculate the p-values with White heteroskedasticity robust covariance estimator. Regarding the volatility panels, as the models are set-up in a GARCH-like fashion, we have a high autocorrelation in the dependent variable. For this reason, we opt for Driscoll-Kraay heteroskedasticity and autocorrelation robust covariance estimator (Driscoll and Kraay, 1998). We shall mention though, that potentially biased estimates due to Nickell-bias[11], should indeed be negligible, as we have a small number of firms but high number of time periods (Nickell, 1981).

[11] Nickell-bias arises in dynamic panel models with fixed effects, due to the demeaning process creating a correlation between the error and regressor.

5 Empirical analysis

In this section, we start by presenting a descriptive statistics analysis. We move forward with a discussion on results relating to the methodology, scoring and aggregation methods. We then present anad analyse the results related to our hypotheses and prior related work. Lastly, we provide a brief discussion around limitations our research.

5.1 Descriptive statistics

We present general descriptive statistics for the variables used in the upcoming regressions for H1, H2 and H3 in Table 5.1. We categorise the descriptive statistics analysis to the order we discuss it: 1) Tweets related statistics and differences in data-sets, 2) Tweets scoring statistics which are the same for both data-sets and 3) Descriptive statistics related to the sentiment-features.

Firstly, we notice that the number of collected tweets for the two data-sets are very different. The total amount of collected tweets for the cashtags data-set is 3.5 million, whereas it is 125.7 million for the general-company data-set. Tables A.1 and A.2 (Appendix) further highlight these differences on a company-level. This is no surprise as one can expect more people are talking about a company in general terms than for its financial performance specifically. It is important to discuss the number of tweets collected, as more tweets should provide more information about the true sentiment of the firm by reducing the impact of outliers, hence making the signal less apt to noise. However, there is also a discussion to be made regarding the connection between public awareness and the number of tweets. Companies with higher public awareness are more likely to draw more public discussion about their affairs, which most certainly affect the number of tweets. Looking at the total number of tweets for companies such as Facebook, Tesla and Walt Disney, it is easy to support this idea. Fewer tweets due to a lack of public interest may consequently lead to a higher share of experts voicing their opinion, which might have a larger impact on investment decisions. This idea supports our second hypothesis, in the sense that the discussion around companies' financial performance is conceived by a higher share of people privy in the matter. This would in turn lead to less noise in the sentiment signal.

We also want to mention the potential implications of the large data loss in the cashtags data-set with our filtering method. The share of tweets filtered differs

Table 5.1: General descriptive statistics

This table presents general descriptive statistics for the variables used in the regressions from both data-sets. The descriptive statistics are calculated from the pooled sample. The number 1, 2 and 3 after sentiment scoring method denotes count-neutral, count-dependent and mean sentiment aggregation method. Under sentiment-features: var denotes daily sentiment variance (calculated as the variance of all tweets' score that day); pos, neg and tot denote number of positive, negative and total tweets each day. All statistics presented are on firm-level, except the S&P500 log return.

	Cashtags				General-company			
Variable	Mean	Std	Max	Min	Mean	Std	Max	Min
Financial data								
Firms log return*	0.0309	1.7521	15.9966	-21.0239				
S&P500 log return*	0.0129	1.0414	4.8403	-4.1843				
Sentiment								
vader_1	0.4199	0.3390	1.0000	-1.0000	0.4181	0.2083	1.0000	-0.6564
vader_2	0.8506	0.7053	5.2095	-2.1972	0.9448	0.5473	4.7381	-1.5709
vader_3	0.1133	0.0865	0.6076	-0.3818	0.1565	0.0831	0.6884	-0.4957
lm_1	-0.2088	0.5099	1.0000	-1.0000	-0.0903	0.3048	1.0000	-1.0000
lm_2	-0.3989	0.8263	3.5835	-3.8501	-0.1971	0.6846	4.4860	-3.7510
lm_3	-0.0638	0.1435	1.0000	-0.8889	-0.0359	0.1171	0.9277	-0.7409
Sentiment-features								
vader_var	0.1024	0.0424	0.4011	0.0000	0.16877	0.03702	0.39984	0.02235
lm_var	0.2454	0.1097	2.0000	0.0000	0.30486	0.06582	0.63126	0.03177
vader_pos	46.71	122.65	2348	0	5958	14129	237024	7
vader_neg	26.12	88.34	1881	0	2351	5398	96287	0
lm_pos	12.67	34.29	717	0	1682	3754	63682	0
lm_neg	24.77	83.52	1924	0	1988	4509	74185	0
tot	121.79	315.15	5823	0	13117	32600	655834	18

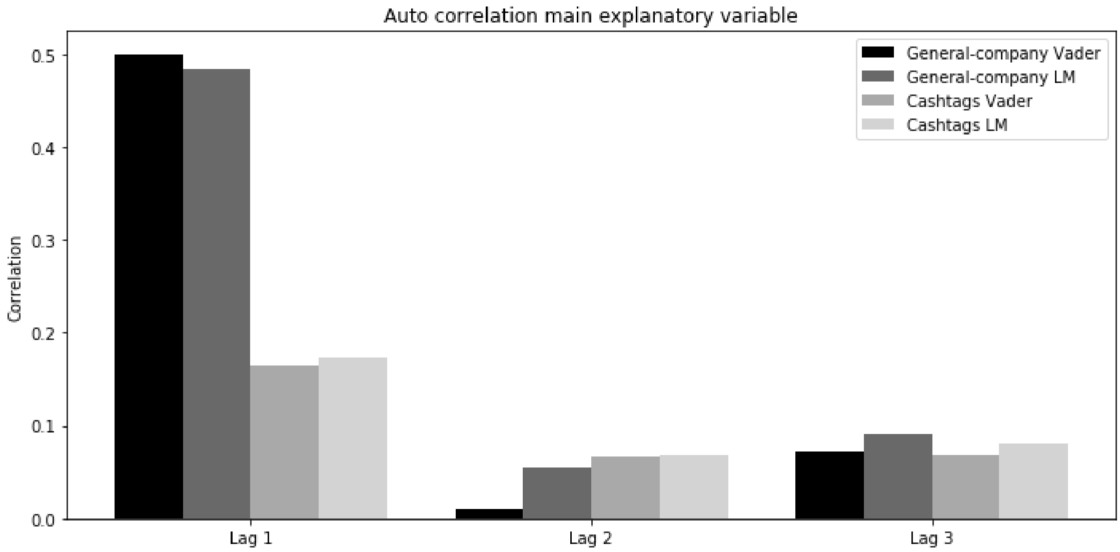
*Note that these are in percentage points and not percent.

a lot between different companies. Berkshire Hathaway, for example, retains less than 10% of its collected tweets while Tesla retains more than half. Filtering out tweets with more than one cashtag is rather strict, and we may lose a lot of relevant sentiment information in the process. However, as neither scoring methods can deduce whether a tweet's sentiment with multiple cashtags is related to firm A or B, we feel it is necessary. When looking at tweets filtered out, we find the vast majority include cashtags for more than five firms. This suggests theat the extra cashtags are included for exposure rather than for conducting a fruitful conversation about the firm's performance.

Outside of filtering out non-English tweets, we do not filter tweets in the general-company data-set. This may cause more noise to be present in regards to sentiment. However, implementing a similar simple filtering method as for the cashtags data-set risk filter out more useful information than noise. We should note that cashtags are very seldom used in the general-company data-set. However, a more sophisticated filtering method should indeed be better at capturing the true sentiment signal in the larger data-set and would probably be deserving of its own paper.

In regards to general statistics, we see some differences in sentiment between the two data-sets, depending on the scoring method. In general, the average sentiment is higher for the general-company data-set when scored by VADER[1], but lower when scored by LM. In addition, we see higher variability in the cashtags data-set, for every scoring and aggregation method. It also has a much higher first-order auto-correlation, as seen in figure 5.1. This is indicative of the general company sentiment being a more long-term sentiment than the cashtags-derived sentiment. This makes sense as the debate regarding company stock tend to be nefariously short-term in the public space.

Figure 5.1: Autocorrelation of the main explanatory variable
Auto correlation for the sentiment scores aggregated with count-neutral aggregation method for VADER and LM on both data-sets, lag 1 to 3.



Secondly, we find interesting differences between the two scoring methods. This is expected as the techniques are very different from each other. LM score more than 70% of tweets as having neutral sentiment, over both data-sets, while VADER only score around 40% as neutral. When we look at the neutral tweets, we find most of them having a score equal to zero. This is especially true for the LM score. A score of exactly zero is indicative of the scoring algorithm not having enough information to produce a score. As VADER is tailored specifically for shorter social media micro-blogging posts, whereas LM is not, it is not difficult to see that LM might be more prone to this problem. However, since LM is tailored for financial texts (albeit longer texts like 10-K filings), the fact that it scores fewer tweets may function as an internal filter in this setting. Tweets with no matching words with the LM financial dictionary are simply given a score of 0. This could, in fact, strengthen the LM method's ability to extract the relevant sentiment signal in regards to the stock, as opposed to VADER which is better at deriving general sentiment from social media posts. Thus, even though VADER outperforms LM in terms of scoring, it may also bring a lot more noise with it. Furthermore, we see that VADER is overwhelmingly

[1]Except for count-neutral VADER, though the difference is minuscule.

positively inclined compared to LM by scoring 64.19% and 71.76% of non-neutral tweets positive as opposed to 33.86% and 45.92%. This may be due to the fact the word list of positive words in LM is somewhat limited. This should not have any overall effect on the results of the upcoming regressions, however but highlights the difference between the two scoring methods. In terms of scoring, we can see that the performance of the sentiment analysis method largely depend on what type of data it was trained on.

Finally, we want to touch briefly on the sentiment-features. As expected, there is a significant difference in magnitude of the number of total, positive and negative tweets between the data-sets. This should not be of concern in the regressions for H3, as we log-transform these. Also, we see the minimum value for the sentiment features on the cashtags data-set is zero. This is due to an inconvenient inconsistency on Twitter, which we rely on for data gathering. On at least one day, for one company, for reasons unbeknownst to us, Twitter reports zero tweets. This observation is treated as a dropped observation. Hence it should not be of concern.

5.2 Results related to methodology

In this section, we focus mainly on results relating to the methodology. We start with a discussion regarding the two scoring method. We find LM perform better than VADER. Then we discuss the different aggregation methods as well as briefly touch on implications of our methods. We find that the count-dependent aggregation method performs the best. In regards to H2 (*Returns are more sensitive to cashtags*), we find that cashtags provide a better sentiment signal. We discuss this more in detail in Section 5.4. Therefore, the discussion regarding the results of H1 (*Twitter sentiment has predictive power on company returns*) in Section 5.3 will be based around count-dependent LM from the cashtags data-set. Lastly, we provide a brief discussion regarding the selected lags.

Tables B.1 and B.2 (Appendix) present all regressions related to H1. By taking an overview of these results, we can see that for the vast majority of specifications, regressions with sentiment derived with LM display a more significant relationship than their VADER counterparts: the R-squared measures are higher, the parameters are more significant and larger. This holds true for both data-sets, although the difference seems larger for cashtags. This is indicative of LM being a better sentiment analysis tool than VADER in terms of extracting sentiment related to asset returns. There are some implications to LM performing better. It implies that the performance of the sentiment analysis method depends more on its lexicon than the type of data it is trained on. In addition, it indicates that predefined treatment of certain sentence characteristics may not play as an important role in this setting.

Finally, the data-sets we use are sure to be filled with noise (especially the larger one). While VADER may be the better sentiment analysis tool for social media in general (Hutto and Gilbert, 2014; Araujo et al., 2016), the fact that it is able to derive sentiment from much more tweets tells us one of two things. Either the sentiment signal is noisier or, the sentiment derived is of a more general character and not as good at modelling asset returns. It is, however, important to mention that for most of the regressions, where the parameters are significant, VADER and LM report the same sign of the coefficients. Thus, we can conclude that they do in fact, model similar relationships. Despite VADER being tailored for social media, we will focus on reporting the results from the LM specifications. For future research however, we believe a more optimal solution would be to combine the two methods, use the general VADER framework, but train it in a financial setting with a financial lexicon.

Similarly, we compare the performance of the three aggregation methods. Again, inspecting Tables B.1 and B.2 (Appendix) we can see that the performance of the three aggregation methods are rather similar: signs of parameters are the same for both control variables and explanatory variables. Significance and explanatory power are in large similar across the methods as well. This is especially true for count-neutral and count-dependent. This is not surprising, as they use the same raw data. Furthermore, we find that the count-dependent aggregation method performs best for the cashtags data-set, providing higher R-squared and more significant parameters than the other two. This result is in line with previous research that finds that inclusion of message volume has a significant impact on results (Antweiler and Frank, 2004). This finding, however, is mostly diminished for the general-company data-set. The differences between count-neutral and count-dependent aggregation are almost non-existent. The differences between the methods in regards to the different specifications are also more inconsistent. For the most part, the methods perform in the order they were presented, count-neutral yielding a very slightly larger R-squared than count-dependent, and mean-sentiment a bit worse. However, in the specifications without contemporaneous explanatory variables, the mean sentiment aggregation method tend to outperform the other two. This result may imply that the mean sentiment aggregation is better at extracting predictive information that carries over to the contemporaneous variables, in this data-set. However, this is not consistent with the results of the cashtags data-set, where count-dependent consistently performs the best, so it may be a spurious result. Touching on the similarities between count-neutral and count-dependent in this data-set, it is most likely related to the logarithmic normalisation, causing the correlation to converge as the number of tweets go to infinity. However, as we find the cashtags data-set generally producing a better sentiment signal (a more in-depth discussion is provided in regards to H2 in Section 5.4), as well as being the preferred choice in the literature (Antweiler and Frank, 2004; Mao et al., 2015), we focus on the results around the count-dependent aggregation.

Lastly, there is no real consensus in the literature as to how many lags to include in regards to modelling returns based on sentiment. Hence, we minimise the Akaike information criterion (AIC) to find the optimal number of sentiment lags in our models. We do this on the specification that includes controls for market returns and the contemporaneous sentiment, for up to four lags, and find that two lags are optimal for us. We use this result for the other specifications, too. Our ambition is to have as few lags as possible, in order to have a parsimonious model. In addition, in line with noise theory, we expect a reversal following an overreaction to happen quickly, as our sample consists of only large firms (Hong and Stein, 1999).

5.3 H1: Twitter sentiment and stock returns

In order to test whether sentiment derived from Twitter can help predict company-level returns, we construct several linear regressions, as discussed in the methods section. We divide this section into two parts. First we present specifications (Table 5.2), with and without the contemporaneous sentiment that do not control for market return. Then we present corresponding specifications with controls for the market (Table 5.3) and discuss key differences.

Table 5.2: H1: Regression results with count-dependent LM

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the cashtags data-set without controlling for market returns. Sent is the count-dependent LM sentiment. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated using Whites HC robust covariance estimator. The stars denote significance level.

		Regressions without control for market return			
#	Independent variable	(1)	(2)	(3)	(4)
1	$sent_t$	0.3307 ****		0.3175 ****	
	$sent_{t-1}$	-0.1245 ****	-0.0559 ***	-0.1436 ****	-0.0672 ****
	$sent_{t-2}$	-0.0854 ****	-0.0442 **		
2	r_{t-1}	-0.0569 ***	-0.0419 **	-0.0518 ***	-0.0384 **
3	Intercept	0.0319 *	0.0290	0.0348 *	0.0315 *
	# obs	8904	9090	9118	9318
	R-Squared	0.0225	0.0032	0.0209	0.0026

Regressions 1 to 4 are presented in the appendix as 37 to 40.
* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

If we start by investigating Table 5.2 column (1) and (3) and specifically at the contemporaneous sentiment variable, we can see significance at 0.1% level. The coefficient is positive, which is to be expected, and in line with the literature (Brown and Cliff, 2004). The parameter value of 0.3307 means that if the sentiment today is 1, we can expect it to rise today's return by 0.3307 percentage points. The literature provides a few reasons why non-lagged explanatory variables show significant relationship with the dependent variable: (i) finer grain effects of the variable may not be visible by the data, (ii) the model is misspecified and variables are related to lags not included in the model, or (iii) the model captures a legitimate effect, although to an extent where it is not possible to model the direction of the relationship (Granger, 1969; Geweke, 1982). The literature suggests that (i) and (iii) are likely in this case. Antweiler and Frank (2004) find predictive power of sentiment on the 15-minute horizon, which support[2] (i). Brown and Cliff (2004) however, find the relationship to be two way directional: sentiment affect returns, but returns also affect sentiment, supporting (iii). However, we want to stress that, since previous research find a predictive sentiment effect on return, we do not investigate this further, rather, we focus on the direction in line with our hypotheses: sentiment predictability over returns. Considering the potential two-way relationship, we cannot reject the null hypothesis of no predictability at this stage.

In contrast to the contemporaneous sentiment variable, we find a significant negative relationship on the lagged sentiment, for all specifications in Table 5.2. Interestingly, this implies a positive sentiment correlates with a negative future return. A possible explanation for this is that the stock reverts back toward its fundamental value after an overreaction, which is in line with the noise theory De Long et al. (1990). The fact that both the lagged return as well as the lagged sentiment show this relationship indicates that not only the return is an overreaction, but that the contemporaneous sentiment experience overreaction, too. This is not surprising considering the two-way relationship the contemporaneous sentiment has with the return. Due to the large companies in our sample, it is not far-fetched to believe that the reversal happens quickly due to efficiencies in the market. It is worth noting though, that lagged sentiment is more significant than the lagged return, which indicates that the sentiment encounter a stronger overreaction.

Another interesting observation is the effect of the contemporaneous sentiment variable. We see the specifications with it have a substantially higher explanatory power compared to the specifications without. Again, this is not surprising considering its relationship with the returns. However, we also find its inclusion increases the lagged sentiment coefficients by more than double, while still being significant at a 0.1% level. Note that this hold for both lags. A possible explanation to this could be that the lagged sentiment signal includes both an uncertain and a more certain element, in addition to noise. Once the contemporaneous sentiment is known, the uncertain element of the lagged sentiment becomes more evident and thus significant in our

[2]Although, as mentioned earlier, not using Twitter sentiment.

estimations.

Table 5.3: H1: Regression results with count-dependent LM, continued

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the cashtags data-set with controls for market returns. Sent is the count-dependent LM sentiment. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated using Whites HC robust covariance estimator. The stars denote significance level.

		Regressions with controls for market return			
#	Independent variable	(5)	(6)	(7)	(8)
1	sent_t	0.2467 ****		0.2427 ****	
	sent_{t-1}	-0.0582 ****	-0.0059	-0.0655 ****	-0.0068
	sent_{t-2}	-0.0329 *	-0.0038		
2	rm_t	1.0412 ****	1.0496 ****	1.0415 ****	1.0487 ****
	rm_{t-1}	0.0404	0.0368	0.0398	0.0362
3	r_{t-1}	-0.0534 **	-0.0405 *	-0.0507 **	-0.0387
4	Intercept	0.0176	0.0139	0.0182	0.0142
	# obs	8904	9090	9118	9318
	R-Squared	0.3969	0.3879	0.3962	0.3872

Regressions 5 to 8 are presented in the appendix as 33 to 36.
* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

In Table 5.3 we present the same specifications adding control for market return. We see some interesting effects. If we look at the models without the contemporaneous sentiment variable, we see that the predictability of sentiment is insignificant on the 10% level. A possible explanation for this is that the companies in the sample are the largest companies in the United States, and make up a substantial share of the S&P500 index[3]. When the market return is included as a control, it absorbs the predictive sentiment signal. Again, however, when the contemporaneous sentiment is included, the sentiment lags become significant again[4]. Furthering the possible explanation, under this theory the predictive sentiment signal might contain both a firm-specific and general market element. When we know the market return today, but not the sentiment, the firm-specific sentiment signal is uncertain. However, when we know today's sentiment, too, the predictive firm-specific signal becomes more clear. This explanation is plausible for two main reasons. Large mature companies tend to move together with the market and people's thoughts and opinions regarding individual companies are to a very large extent in line with their opinion of the general market. When the market is nearing the end of a cycle, people become less optimistic about individual companies' performance as well.

We conclude that sentiment on Twitter indeed can provide valuable information and

[3]The 30 largest companies of the S&P500 make up around 40% of the weight.

[4]Although, the second sentiment lag a bit less.

help both explain and, to some extent, predict individual companies' stock return. The small R-square for the predictive regressions, however, is a testament to how difficult and complicated it is to predict daily stock market returns. That being said, even a small increase of 0.2%[5] predictability on daily returns may give an investor a useful edge.

5.4 H2: Cashtags and general-company tweets

In order to test whether stock movements are more sensitive to sentiment derived from cashtags than from the general-company data-set, we construct additional regressions, including sentiment from both data-sets in each regression. We structure this section into two parts. First we present the results of select regressions with sentiment from both data-sets (Table 5.4), and then we present a comparison of corresponding specifications with sentiment from each data-set separately (Table 5.5).

If we investigate Table 5.4 columns (1) and (2), we see that the contemporaneous sentiment variable from both data-sets is positive and significant at the 0.1% level. Following the discussion presented in the previous section, this may be due to a two-way relationship with returns. However, it also suggests that the two sentiment metrics model similar relationships. The coefficient is larger for the general-company sentiment in both regressions, which indicate that the relationship is stronger. However, if we focus on the lagged sentiment metrics, we find that the significance of the general-company sentiment is lower than its cashtags counterpart. This is especially true in regression 3, when the contemporaneous variables are omitted. This might tell us that the stronger relationship for the contemporaneous sentiment is linked with the two-way relationship, rather than carrying any significantly larger predictive signal.

However, we also see that coefficient for the lag one sentiment is almost twice as large for general-company than cashtags when the contemporaneous variable is included. If we follow the idea from the previous section regarding the sentiment signal consisting of both a firm-specific and general market elements, one can argue that the firm-specific signal is stronger, yet more uncertain for the general-company sentiment. It's worth noting though, that the higher autocorrelation of the general-company sentiment, as seen in figure 5.1, may contribute significantly to this effect.

When we instead estimate the same regressions with the two sentiments separately,

[5]Note that this is compared to estimating an AR(1) model on the returns (that is, the same model as Table 5.2 column 2 with sentiment terms omitted). We do not present the results of the AR(1) estimation, however they can be provided upon request.

Table 5.4: H2: Combined sentiment regression results with count-dependent LM

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from both data-sets. Cash_sent is the LM sentiment derived from the cashtags data-set. Gc_sent is the LM sentiment derived from the general-company data-set. Rm is the S&P500 index log return. R is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated using Whites HC robust covariance estimator. The stars denote significance level.

		Regressions with combined sentiment		
#	Independent variable	(1)	(2)	(3)
1	cash_sent_t	0.2105 ****	0.2864 ****	
	cash_sent_{t-1}	-0.0581 ****	-0.1165 ****	-0.0574 ***
	cash_sent_{t-2}	-0.0270	-0.0770 ****	-0.0447 *
2	gc_sent_t	0.2987 ****	0.3493 ****	
	gc_sent_{t-1}	-0.1196 **	-0.2214 ***	-0.0308
	gc_sent_{t-2}	-0.0414	-0.0127	0.0322
3	rm_t	1.0297 ****		
	rm_{t-1}	0.0493 *		
4	r_{t-1}	-0.0586 **	-0.0573 ***	-0.0429 **
5	Intercept	0.0181	0.0320 *	0.0286
	# obs	9300	9300	9300
	R-squared	0.3955	0.0278	0.0035

Regressions 1 to 3 are presented in the appendix as 129, 133 and 134.
* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

as seen in Table 5.5, we find more evidence that we can reject the null hypothesis of there not existing a significant difference in sensitivity to sentiment between the data-sets. If we compare (4) and (6) we see similar results as in the corresponding combined estimation (1). The contemporaneous sentiment is positive and significant at the 0.1% level for both sentiments, although with a greater coefficient for the general-company. In addition, we see that both of the sentiment lags are strictly more significant and that the explanatory power is strictly larger for the cashtags-derived sentiment. This further support our second hypothesis. Moreover, these results support the discussion regarding possible differences between the sentiment in the two data-sets mentioned in section 5.1. If the larger data-set contain more general opinions and long-term sentiment about the companies, than about their financial performance, then we can expect cashtags-derived sentiment to show more significant signs of a predicting relationship with returns, which is what the evidence suggest indeed. Under this assumption, the hypothesis boils down to whether the sentiment related to the firms' financial performance is a better predictor of short term returns than the more general sentiment. This is consistent with noise theory.

Table 5.5: H2: Comparison of regression results with count-dependent LM

In this table we present a comparison of the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the cashtags or the general-company data-set. Sent is the count-dependent LM sentiment from either cashtags or general-company data-set, as denoted in the column header. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated using Whites HC robust covariance estimator. The stars denote significance level.

#	Independent variable	Cashtags sentiment (4)		Cashtags sentiment (5)		General-company sentiment (6)		General-company sentiment (7)	
1	sent_t	0.2467	****			0.3314	****		
	sent_{t-1}	-0.0582	****	-0.0559	***	-0.1280	**	-0.0446	
	sent_{t-2}	-0.0329	*	-0.0442	**	-0.0321		0.0143	
2	rm_t	1.0412	****			1.0322	****		
	rm_{t-1}	0.0404				0.0478	*		
3	r_{t-1}	-0.0534	**	-0.0419	**	-0.0473	**	-0.0427	**
4	Intercept	0.0176		0.0290		0.0169		0.0315	*
	# obs	8904		9090		9300		9300	
	R-Squared	0.3969		0.0032		0.3899		0.0020	

Regressions 4 to 7 are presented in the appendix as 33, 38, 81 and 86.
* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

The general sentiment may indeed have an effect on the stocks, but it is not hard to believe such an effect will take longer than a day to materialise.

Another possible explanation to cashtags being the better performer may be that our sentiment analysis tools are not able to capture the sentiment information set well enough in the larger data-set. Bollen et al. (2011) find that some 'moods', other than just a positive or negative sentiment, possess predictive capabilities in regards to stock index returns. It is plausible that the larger data-set contains more multi-faceted sentiment than what our sentiment analysis tools can extract. In that case, sentiment that neither fall into the positive or negative category, but rather some other 'mood', may be misclassified and hence contribute to more noise. This explanation is also consistent with Zhang et al.'s (2011) finding that certain classes of words to wield predictive power over returns.

In conclusion, we find that sentiment derived from cashtags is better able to explain the variability in returns than the general-company counterpart, and that it possesses better predicting capabilities. The general-company data-set may, however, still contain more sentiment information, but in a multi-faceted form that our uni-dimensional sentiment analysis tools are unable to extract.

5.5 H3: Twitter sentiment and stock volatility

In order for us to test our third hypothesis, we estimate several different regressions on the estimated conditional volatility that we obtained from GARCH with select sentimen-features. In Table 5.6 we present the results of the different regressions.

Table 5.6: H3: Volatility regression results with count-dependent LM

In this table we present the result of the regressions that try to model the relationship between company-level conditional volatility and select sentiment-features derived from the cashtags data-set. Sent^2 is the squared count-dependent LM sentiment. #total is the total number of tweets. #pos and #neg are the number of positive and negative tweets. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated using Whites HC robust covariance estimator. The stars denote significance level.

		GARCH-like regressions for volatility					
#	Independent variable	(1)	(2)	(3)	(4)	(5)	(6)
1	sent^2$_{t-1}$	-0.0001	-0.0004				
2	#total$_{t-1}$	-0.0051		0.0119 ****			
3	#pos$_{t-1}$	0.0110 ****			0.0094 ****		
4	#neg$_{t-1}$	0.0099 ****				0.0082 ****	
5	sent_var$_{t-1}$	-0.0533 ****					0.0034
6	resid$_{t-1}$	-0.4373 **	-0.4373 **	-0.4389 **	-0.4611 **	-0.4177 **	-0.4386 **
7	cond_vol$_{t-1}$	0.7999 ****	0.8169 ****	0.8029 ****	0.8077 ****	0.8061 ****	0.8168 ****
8	Intercept	0.0480 ****	0.0479	0.0083	0.0356 ****	0.0349 ****	0.0466 ****
	# obs	9078	9108	9300	9300	9300	9324
	R-squared	0.6871	0.6760	0.6864	0.6844	0.6832	0.6769

Regressions 1 to 6 are presented in the appendix as 153, 154, 157, 161, 162 and 163.
* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table 5.6 shows that the lagged sentiment does not significantly explain one day ahead conditional volatility, in either estimation (1) nor (2). Moreover, we see that the total number of tweets have a significant positive effect on the conditional volatility in estimation (3). This is in line with the results that Antweiler and Frank (2004) find on message postings on Yahoo!Finance and Ragingbull. A possible explanation which previous research has shown is that the number of tweets correlates with the trading volume and that the trading volume has a significant effect on the volatility. Intuitively, more postings on Twitter should increase the available information which, according to Danthine and Moresi (1993) would decrease the volatility; however, our results suggest the opposite. Furthermore, we see that the effect is no longer significant when we include the other variables in estimation (3). As the total number of tweets are closely related to the number of positive and negative tweets, the relationship may in (1) instead be captured by the positive and negative counts.

If we further investigate the number of positive and negative tweets, we find that both variables are positive and significantly explains one day ahead conditional volatility. This relationship holds both when regressed on their own (4) and (5), as well as in the kitchen-sink estimation with all sentiment-features included (1). One interesting observation is that the estimations suggest that positive tweets increase the volatility more than negative tweets do. A possible explanation for this may be that when the mood is positive investors trade more, while people are more cautious when the mood is negative.

Furthermore, we find in estimation (1) that the variance in daily sentiment has a significant negative relation with volatility. As our proxy for disagreement of investors, previous studies show inconclusive results in its effect on the volatility. For example, Sprenger et al. (2014) find disagreement to have a positive relationship with volatility while Antweiler and Frank's (2004) results are in line with our own. Even though financial theory suggests that disagreement should induce trading and volatility since the assets are valued differently (Harris and Raviv, 1993). Our results could be explained by that the disagreement between investors may release additional information due to extensive debating, as some previous studies suggest (Danthine and Moresi, 1993; Das et al., 2005). In turn, this helps rational investors counteract the actions of noise traders, thus decreasing volatility. Another possible explanation for the results is that disagreement decreases the amount of trading due to investors being risk-averse, and aware of the fact that the other party will only enter a trade if they have an advantage. It is important to note that in estimation (6), when we test the sentiment variance by itself, it does show a significant relationship, which might indicate that the significant result in (1) is spurious.

To conclude, we see that Twitter data can help to explain stock volatility. The most significant and robust variables seem to be the amount of positive and negative tweets. We find no evidence that the sentiment have any significant explanatory power for the volatility even though number of positive and negative tweets can be seen as a sort of sentiment. Our result does not support previously found relationship between the total number of tweets and volatility. Furthermore, we find, in line with some previous literature, that disagreement in the sentiment decreases the stock volatility (Antweiler and Frank, 2004).

5.6 Summary of results

To conclude, we find evidence that supports previous research and add unique findings to the field by testing whether more general company tweets actually provides better explanatory power than cashtag-tweets. Firstly, we see that sentiment expressed on Twitter can be used to predict stock return. The result shows that the first and second sentiment lag can help explain the stock return, and that the relationship is negative. This is in line with previous research that find the same relationship (Yu et al., 2013).

The noise trade theory provides a possible explanation to why positivity in the public mood may lead to a negative return in the following days. The theory suggests that asset prices tend to revert to their fundamental values within a week from the overreaction. This happens faster for companies that are bigger because the pricing is more efficient. In conjunction with the same day sentiment being significantly positively correlated, this may indicate that people tend to overreact on the 'first day' and then the prices start reverting back toward fundamentals as quickly as the next day. This is consistent with noise theory as we only have the largest companies in our sample (Hong and Stein, 1999).

We also test whether advances in social-media sentiment analysis translates into better performance in terms of financial modelling. We find that the finance-focused LM-method is still better at extracting the relevant sentiment for financial modelling. Furthermore, we find very little difference in the performance of the two sentiment aggregation methods suggested by Antweiler and Frank (2004). Contrary to what Smailović et al. (2013) suggests, we find models than include neutral sentiment tweets consistently performs worse than those not omitting neutral tweets.

Secondly, our study concludes the sentiment derived from cashtag-tweets can better explain and predict company-level returns than sentiment derived from more general company tweets. We provide two possible explanations for this result. (i) Cashtag-tweets contain sentiment regarding the firm's financial performance while the general company tweets contain more general opinions and long-term sentiment. (ii) There is more noise in the more general tweets and that our sentiment analysis tools are not sophisticated enough to extract what is relevant.

Lastly, we examine if Twitter data can help explain stock market volatility. The results suggest that the sentiment has no significant effect on the volatility. However, we find that variance in the sentiment has a negative significant effect, which is in line with empirical findings of Danthine and Moresi (1993). These results are not consistent with findings of disagreement between investors increases volatility. We also find some evidence that the number of tweets have a significant effect on stock market volatility. When there is an increase in user postings on Twitter volatility increases. While this might be a two-way relationship, we do not explore this in our paper.

5.7 Discussion and limitations

We want to end this part with a small discussion surrounding the limitation of our study. It is important to comment on our time-frame and what effects it might have on our results and our generalisation to other periods. We use a time frame of 15 months, stretching from 2017-12-31 to 2019-03-31, which is longer than than any previous research on sentiment and daily company-level returns. The year 2018 was the worst performing year since the financial crisis in 2008 for the S&P500 returns. This may have influenced our results and the ability to be generalised for other periods. Although, the beginning of 2019 performed very well in contrast, negating any such effect. Another limitation with our data is that we use daily data. As previously discussed, the finer grain effects might not be visible for daily data. Investigation intraday data may help capture some of the forecasting effects that we do not capture as it may have already reverted in our daily data-set.

Bollen et al. (2011) find that linear regression is not the preferred way to model the sentiment effect on stock returns. The authors instead show that a non-linear model can help increase the significance of the relationship between public sentiment and market returns. Given this, it is possible that our linear regressions could perform better with some non-linear transformations.

Lastly, several of the previous studies in the field have used other polarity and mood dimensions. As we choose to only divide our sentiment into negative and positive, we lose out on information about other sentiment dimensions that might provide additional predictability. For example, Bollen et al. (2011) find that it may be better to have more sentiment dimensions (they use a total of six moods) when explaining the stock returns and find that 'calm' can predict stock return.

6 Conclusion

In this paper we examine if the sentiment on Twitter can help predict company-level stock movements. In relation to previous research we provide the most comprehensive study to date in regards to number of tweets, time frame and companies. We compare and test two different sentiment analysis methods, and find that one consistently outperforms the other. We also contribute to the discussion surrounding sentiment-signals from cashtag-tweets and broader company tweets.

The main conclusions from our results are that the expressed sentiment on Twitter provide valuable information when predicting stock returns. We show that there is a significant negative relationship between sentiment variables lagged one and two periods and company-level returns. We argue that the negative nature of the relationship is due to the empirical finding that stock prices revert to the fundamentals shortly after an overreaction. This is in line with prior related work that have shown similar results when predicting market indices returns (Mao et al., 2015). In addition to this, we find the finance-focused sentiment analysis method of LM to outperform the results of social-media tailored Vader. We draw the conclusion that a financial lexicon is better suited for predicting stock returns using sentiment analysis, even on medium where it was not initially intended.

With our large data-set we are also able to compare tweets containing cashtags with tweets containing the general company name. Previous research suggest that the additional information that is provided by more tweets should give a more accurate estimation of the sentiment, although this might be the case we find that the smaller data-set consistent of cashtag-tweets has a better forecasting predictability for stock return. We draw the conclusion that there is much more noise in the larger data set for the short term predictability of stock returns. Lastly we see that Twitter also can help explain some stock market effects. We find evidence that support and confirms previous research that variance in the sentiment can help explain volatility and that the total amount of tweets has a positive correlation with volatility. We show that when controlling for contemporaneous sentiment, the predicting power of lagged sentiment strengthens, while controlling for market returns weakens it. We provide an explanation for this in that the Twitter sentiment contains both a firm-specific and a general market element. However, this is ground for future research.

Due to Twitter's ability to capture important information regarding forecasting of stock returns we believe our paper holds a practical importance for the construction of future high-frequency investment strategies. However, we realise that there is much more research to be done in this field and therefore encourage others to continue contributing to the research field.

Several recommendations for future research can be suggested. In regards to the methodology, exploring more dimensions in sentiment than just positive-negative is interesting. As could expanding VADER with a financial lexicon, in order to capture more sophisticated sentence characteristics in a finance setting. Investigating effectual forecasting on a longer term is also of interest and could hence be considered. In terms of data collection, a longer time-frame, in order to capture a full cycle, rather than just part of it may also give more insight in the twitter and stock movement relationship. It would also be interesting to test if there are differences in the relationship between different timings in the cycle, i.e. market contraction compared to expansion. Moreover, testing non-linear models for the relationship can be an interesting contribution.

Bibliography

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

M. Araujo, J. Reis, A. Pereira, and F. Benevenuto. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1140–1145. ACM, 2016.

S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society, 2010.

N. Barberis, A. Shleifer, and R. Vishny. A model of investor sentiment. *Journal of financial economics*, 49(3):307–343, 1998.

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

G. W. Brown and M. T. Cliff. Investor sentiment and the near-term stock market. *Journal of empirical finance*, 11(1):1–27, 2004.

K. C. Butler and S. J. Malaikah. Efficiency and inefficiency in thinly traded stock markets: Kuwait and saudi arabia. *Journal of Banking & Finance*, 16(1):197–210, 1992.

R. Chen and M. Lazer. Sentiment analysis of twitter feeds for the prediction of stock market movement. *stanford edu Retrieved January*, 25:2013, 2013.

P. H. Cootner. The random character of stock market prices. 1964.

K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, 2017.

M. Daniel, R. F. Neves, and N. Horta. Company event popularity for financial markets using twitter and sentiment analysis. *Expert Systems with Applications*, 71:111–124, 2017.

J.-P. Danthine and S. Moresi. Volatility, information and noise trading. *European Economic Review*, 37(5):961–982, 1993.

S. Das, A. Martínez-Jerez, and P. Tufano. einformation: A clinical study of investor discussion and sentiment. *Financial Management*, 34(3):103–137, 2005.

S. R. Das and M. Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1375–1388, 2007.

J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738, 1990.

E. Demers, C. Vega, et al. *Soft information in earnings announcements: News or noise?*, volume 951. Board of Governors of the Federal Reserve System, 2008.

J. C. Driscoll and A. C. Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560, 1998.

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

E. F. Fama. The behavior of stock-market prices. *The journal of Business*, 38(1):34–105, 1965.

E. F. Fama. Efficient capital markets: Ii. *The journal of finance*, 46(5):1575–1617, 1991.

J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313, 1982.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

D. Griffin and A. Tversky. The weighing of evidence and the determinants of confidence. *Cognitive psychology*, 24(3):411–435, 1992.

M. Harris and A. Raviv. Differences of opinion make a horse race. *The Review of Financial Studies*, 6(3):473–506, 1993.

Harvard. General inquirer dictionaries. Web page, 2019. URL `http://www.wjh.harvard.edu/~inquirer/homecat.htm`. Accessed 2019/04/23.

D. Hirshleifer and T. Shumway. Good day sunshine: Stock returns and the weather. *The Journal of Finance*, 58(3):1009–1032, 2003.

H. Hong and J. C. Stein. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance*, 54(6):2143–2184, 1999.

C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

J. M. Karpoff. A theory of trading volume. *The Journal of Finance*, 41(5):1069–1087, 1986.

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

X. Li, H. Xie, L. Chen, J. Wang, and X. Deng. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23, 2014.

B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

A. W. Lo and A. C. MacKinlay. *A non-random walk down Wall Street*. Princeton University Press, 2002.

T. Loughran and B. McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

T. Loughran and B. McDonald. Textual analysis resources. Web page, 2019. URL `https://sraf.nd.edu/textual-analysis/resources/`. Accessed 2019/04/25.

H. Mao, S. Counts, and J. Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.

H. Mao, S. Counts, and J. Bollen. Quantifying the effects of online bullishness on international financial markets. Technical report, ECB Statistics Paper, 2015.

Y. Mao, W. Wei, B. Wang, and B. Liu. Correlating s&p 500 stocks with twitter data. In *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*, pages 69–72. ACM, 2012.

P. Milgrom and N. Stokey. Information, trade and common knowledge. *Journal of economic theory*, 26(1):17–27, 1982.

A. Mittal and A. Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

S. Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426, 1981.

J. R. Nofsinger. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3): 144–160, 2005.

N. Oliveira, P. Cortez, and N. Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 31. ACM, 2013.

V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE, 2016.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

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

B. Qian and K. Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, 2007.

T. Rao and S. Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)*, pages 119–123. IEEE Computer Society, 2012.

F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23, 2016.

E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM, 2012.

S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88. Springer, 2013.

T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welpe. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.

P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

B. Twedt and L. Rees. Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy*, 31(1):1–21, 2012.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

Yahoo!Finance. Online Database, 2019. Accessed 2019/04/25.

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

X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

A Companies and Twitter statistics

Table A.1: Cashtags: Twitter and firms data table

In this table we present the firms and information regarding their corresponding tweets. We present the search query terms used to collect each firm's tweets, in this case each company's cashtag. We also present the positivity of tweets, defined as the share of positive tweets for each company, excluding neutral tweets. We also present the share of neutral tweets for each scoring method. In total, after filtering, this corresponds to approximately 1.2 million tweets in this data-set.

		Twitter data			Positivity		Neutrality	
	Company	Search query	Collected Tweets	After Filtering	Vader	LM	Vader	LM
1	Alphabet	$goog, $googl	184,205	43,294	68.15%	37.38%	52.86%	76.87%
2	Amazon	$amzn	467,732	148,774	70.97%	40.26%	36.86%	72.45%
3	Apple	$aapl	422,356	136,248	62.68%	36.46%	46.54%	71.92%
4	AT&T	$t	60,099	15,211	64.41%	30.02%	48.29%	74.70%
5	Bank Of America	$bac	90,163	17,076	72.05%	37.77%	48.14%	78.09%
6	Berkshire Hathaway	$brk, $brk.a, $brk.b	63,538	4,219	77.00%	55.96%	44.87%	72.17%
7	Boeing	$ba	84,822	34,301	63.05%	32.57%	44.34%	70.31%
8	Chevron	$cvx	27,444	6,458	76.27%	37.07%	47.74%	76.82%
9	Cisco	$csco	38,853	11,360	82.55%	52.21%	49.45%	75.47%
10	Citigroup	$c	50,209	10,255	68.19%	35.66%	48.65%	74.68%
11	Coca-Cola	$ko	29,682	8,589	75.77%	41.71%	48.29%	79.01%
12	ExxonMobil	$xom	46,342	10,859	72.26%	32.98%	49.07%	78.06%
13	Facebook	$fb	349,013	137,603	58.78%	28.97%	53.57%	75.02%
14	Home Depot	$hd	34,808	10,235	73.32%	43.03%	45.77%	73.45%
15	Intel	$intc	72,221	17,904	69.80%	33.17%	48.75%	72.11%
16	J.P. Morgan	$jpm	88,148	19,962	69.51%	35.17%	53.67%	75.59%
17	Johnson&Johnson	$jnj	52,439	14,010	60.72%	28.09%	43.73%	67.99%
18	Mastercard	$ma	27,312	6,716	77.80%	46.01%	49.23%	81.72%
19	Merck	$mrk	38,788	10,411	74.41%	54.25%	44.67%	76.38%
20	Microsoft	$msft	118,399	24,471	75.06%	48.37%	48.37%	75.33%
21	Nvidia	$nvda	124,864	32,171	66.75%	42.81%	40.23%	69.01%
22	Oracle	$orcl	33,975	9,494	73.45%	36.62%	52.95%	79.38%
23	Pfizer	$pfe	39,476	8,165	72.61%	38.27%	45.67%	72.12%
24	Proctor & Gamble	$pg	32,671	9,082	75.38%	39.89%	46.03%	75.02%
25	Tesla	$tsla	618,493	362,495	57.87%	29.08%	32.08%	61.91%
26	UnitedHealth	$unh	25,448	7,255	76.86%	41.38%	47.82%	78.88%
27	Verizon	$vz	37,997	10,323	74.04%	40.05%	46.49%	74.24%
28	Visa	$v	82,244	42,313	62.23%	35.65%	60.91%	80.88%
29	Wal-Mart	$wmt	59,566	17,630	71.17%	39.00%	46.93%	73.70%
30	Walt Disney	$dis	67,283	23,453	70.69%	44.64%	43.83%	72.24%
31	Wells Fargo	$wfc	60,538	17,011	78.96%	20.72%	27.54%	71.66%
	Total		**3,529,128**	**1,227,348**	**453,946**	**122,994**	**520,168**	**864,103**
	Average		**113,843**	**39,592**	**64.19%**	**33.86%**	**42.38%**	**70.40%**

Table A.2: General-company: Twitter and firms data table

In this table we present the firms and information regarding their corresponding tweets. We present the search query terms used to collect each firm's tweets*. We also present the positivity of tweets, defined as the share of positive tweets for each company, excluding neutral tweets. We also present the share of neutral tweets for each scoring method. In total, this corresponds to approximately 126 million tweets in this data-set.

		Tweet Data		Positivity		Neutrality	
#	Firm	Search Query	Collected tweets	VADER	LM	VADER	LM
1	Alphabet	google	22,770,900	70.48%	42.89%	35.06%	72.14%
2	Amazon	amazon	20,938,630	82.90%	59.19%	12.93%	68.65%
3	AT&T	at&t	1,257,303	62.03%	35.92%	40.88%	69.25%
4	Bank of America	bofa, bank of america, bankofamerica	571,968	64.01%	32.88%	30.82%	68.03%
5	Berkshire Hathaway	berkshire hathaway, berkshirehathaway, berkshire-hathaway	127,443	85.46%	57.88%	41.22%	70.21%
6	Boeing	boeing	1,175,380	55.75%	25.81%	39.81%	65.51%
7	Chevron	chevron	247,963	71.22%	44.20%	41.88%	74.43%
8	Cisco	cisco	1,026,288	84.82%	61.38%	31.47%	66.51%
9	Citigroup	citi, citigroup, citibank	522,112	67.27%	38.49%	35.78%	68.18%
10	Coca-Cola	coca cola, cocacola, coca-cola	982,113	74.71%	57.90%	39.86%	73.82%
11	Walt Disney	walt disney, waltdisney, walt-disney, disney	9,799,637	71.99%	57.77%	33.05%	70.95%
12	ExxonMobil	exxon, exxonmobil, exxon mobil, exxon-mobil	284,717	58.87%	24.01%	31.38%	59.62%
13	Facebook	facebook	37,665,241	66.05%	37.87%	54.94%	79.62%
14	Home Depot	homedepot, home depot, home-depot	649,906	68.98%	50.71%	35.22%	68.72%
15	Intel	intel	2,735,276	58.71%	27.57%	32.74%	56.57%
16	J.P. Morgan	jpm, jpmorgan, jp morgan, j p morgan, j.p.morgan, j.p. morgan	600,007	71.94%	32.40%	31.78%	67.15%
17	Johnson&Johnson	johnson&johnson, johnsonjohnson, johnson & johnson, jnj, johnson johnson	245,664	64.91%	34.11%	37.64%	68.03%
18	Mastercard	mastercard	383,393	77.44%	56.63%	32.44%	68.36%
19	Merck	merck	125,033	65.25%	48.14%	31.32%	61.42%
20	Microsoft	microsoft	6,519,179	75.09%	51.96%	37.09%	70.04%
21	Nvidia	nvidia	849,103	78.21%	67.88%	48.57%	71.75%
22	Oracle	oracle	1,267,920	77.81%	54.10%	36.28%	70.03%
23	Pfizer	pfizer	160,657	59.62%	34.96%	32.26%	60.67%
24	Procter & Gamble	procter gamble, proctergamble, procter&gamble, p&g, procter & gamble, procter-gamble	418,389	73.46%	46.95%	25.34%	64.02%
25	Tesla	tesla	4,545,897	64.55%	38.95%	34.95%	66.70%
26	UnitedHealth	unitedhealth	38,848	82.23%	34.80%	42.62%	69.75%
27	Verizon	verizon	1,199,189	64.89%	47.96%	31.23%	62.63%
28	Visa	visa	3,103,074	64.60%	40.88%	23.59%	53.19%
29	Wal-Mart	walmart, wal-mart, wal mart	4,813,329	63.01%	42.52%	39.93%	74.79%
30	Wells Fargo	wells fargo, wells-fargo, wellsfargo	700,227	69.80%	24.54%	15.20%	61.42%
	Total		**125,724,786**	**56,718,354**	**16,037,228**	**46,686,336**	**90,802,101**
	Average		**4,190,826**	**71.76%**	**45.92%**	**37.13%**	**72.22%**

*Note that for Alphabet we search for 'google' instead, as alphabet in itself is an inherently generic term, in conjunction with it publicly being referred to its old name, Google.

B Regression results

In this appendix we present the results of all regressions we estimate, for all combinations of scoring and aggregation methods. Note that the presentation of the regressions are transposed (i.e. each row respresents a regression, and each column an explanatory variable) in order to condense and save space, in contrast to the conventional way of presenting regression results.

Table B.1: H1: Regression result for cashtag tweets data-set

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the cashtags data-set. The score column denote which sentiment analysis tool is used to derive sentiment. The agg column denote which aggregation method is used to construct sentiment time-series: 1, 2 and 3 denote count-neutral, count-dependent and mean-sentiment as defined in Equations 4.2, 4.3 and 4.4. Sent is the sentiment. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated with Whites HC robust covariance estimator. The stars denote significance level.

reg #	Score	Agg	Intercept	r_lag_1	rm	rm_lag_1	sent	sent_lag_1	sent_lag_2	R-Squared	# obs
1	Vader	1	0.0171	-0.0489 **	1.0409 ****	0.0411	0.3883 ****	-0.0388	-0.0227	0.3940	9511
2	Vader	1	0.0165	-0.0397 *	1.0454 ****	0.0376		0.0240	0.0094	0.3891	9511
3	Vader	1	0.0166	-0.0481 **	1.0407 ****	0.0401	0.3898 ****	-0.0427		0.3931	9541
4	Vader	1	0.0161	-0.0390	1.0451 ****	0.0363		0.0265		0.3881	9541
5	Vader	1	0.0326 *	-0.0514 ***			0.5173 ****	-0.1558 ***	-0.0970 **	0.0108	9511
6	Vader	1	0.0318 *	-0.0410 **				-0.0723	-0.0543	0.0021	9511
7	Vader	1	0.0342 *	-0.0499 ***			0.5134 ****	-0.1686 ****		0.0105	9541
8	Vader	1	0.0336 *	-0.0398 **				-0.0777		0.0019	9541
9	Vader	2	0.0175	-0.0510 **	1.0404 ****	0.0404	0.2299 ****	-0.0383 **	-0.0253	0.3963	9511
10	Vader	2	0.0165	-0.0397 *	1.0453 ****	0.0375		0.0095	-0.0003	0.3891	9511
11	Vader	2	0.0169	-0.0499 **	1.0405 ****	0.0395	0.2289 ****	-0.0440 **		0.3953	9541
12	Vader	2	0.0161	-0.0389	1.0450 ****	0.0364		0.0098		0.3881	9541
13	Vader	2	0.0329 *	-0.0536 ***			0.2863 ****	-0.0932 ****	-0.0653 ***	0.0133	9511
14	Vader	2	0.0318 *	-0.0410 **				-0.0338	-0.0344	0.0022	9511
15	Vader	2	0.0346 *	-0.0515 ***			0.2811 ****	-0.1049 ****		0.0128	9541
16	Vader	2	0.0336 *	-0.0396 **				-0.0389 *		0.0019	9541
17	Vader	3	0.0175	-0.0492 **	1.0402 ****	0.0399	1.7355 ****	-0.2338 *	-0.2330	0.3950	9506
18	Vader	3	0.0167	-0.0397 *	1.0448 ****	0.0375		0.0722	-0.0809	0.3888	9507
19	Vader	3	0.0168	-0.0479 **	1.0405 ****	0.0388	1.7297 ****	-0.2911 **		0.3943	9537
20	Vader	3	0.0161	-0.0387	1.0449 ****	0.0363		0.0448		0.3881	9538
21	Vader	3	0.0330 *	-0.0521 ***			2.2017 ****	-0.6527 ****	-0.4537 **	0.0121	9506
22	Vader	3	0.0322 *	-0.0411 **				-0.2665	-0.2633	0.0021	9507
23	Vader	3	0.0342 *	-0.0502 ***			2.1770 ****	-0.7344 ****		0.0117	9537
24	Vader	3	0.0336 *	-0.0396 **				-0.3139 *		0.0019	9538
25	LM	1	0.0174	-0.0487 **	1.0461 ****	0.0392	0.2980 ****	-0.0686 ***	-0.0300	0.3924	8904
26	LM	1	0.0139	-0.0405 *	1.0496 ****	0.0369		-0.0135	-0.0014	0.3879	9090
27	LM	1	0.0180	-0.0464 *	1.0460 ****	0.0383	0.2921 ****	-0.0742 ***		0.3917	9118
28	LM	1	0.0142	-0.0386	1.0487 ****	0.0362		-0.0153		0.3872	9318
29	LM	1	0.0318 *	-0.0518 ***			0.3900 ****	-0.1587 ****	-0.1055 ***	0.0136	8904
30	LM	1	0.0290	-0.0426 **				-0.0876 ***	-0.0632 *	0.0030	9090
31	LM	1	0.0347 *	-0.0477 **			0.3709 ****	-0.1791 ****		0.0125	9118
32	LM	1	0.0315 *	-0.0395 **				-0.1026 ****		0.0026	9318
33	LM	2	0.0176	-0.0534 **	1.0412 ****	0.0404	0.2467 ****	-0.0582 ****	-0.0329 *	0.3969	8904
34	LM	2	0.0139	-0.0405 *	1.0496 ****	0.0368		-0.0059	-0.0038	0.3879	9090
35	LM	2	0.0182	-0.0507 **	1.0415 ****	0.0398	0.2427 ****	-0.0655 ****		0.3962	9118
36	LM	2	0.0142	-0.0387	1.0487 ****	0.0362		-0.0068		0.3872	9318
37	LM	2	0.0319 *	-0.0569 ***			0.3307 ****	-0.1245 ****	-0.0854 ****	0.0225	8904
38	LM	2	0.0290	-0.0419 **				-0.0559 ***	-0.0442 **	0.0032	9090
39	LM	2	0.0348 *	-0.0518 ***			0.3175 ****	-0.1436 ****		0.0209	9118
40	LM	2	0.0315 *	-0.0384 **				-0.0672 ****		0.0026	9318
41	LM	3	0.0174	-0.0498 **	1.0391 ****	0.0421	1.1308 ****	-0.1660 *	-0.1651 *	0.3961	9506
42	LM	3	0.0167	-0.0400 *	1.0450 ****	0.0377		0.0623	-0.0196	0.3888	9507
43	LM	3	0.0166	-0.0483 **	1.0396 ****	0.0411	1.1147 ****	-0.2004 **		0.3953	9537
44	LM	3	0.0161	-0.0390	1.0451 ****	0.0365		0.0602		0.3881	9538
45	LM	3	0.0330 *	-0.0516 ***			1.4641 ****	-0.4820 ****	-0.3871 ****	0.0145	9506
46	LM	3	0.0323 *	-0.0410 **				-0.1876 *	-0.2004 *	0.0023	9507
47	LM	3	0.0340 *	-0.0489 **			1.4208 ****	-0.5648 ****		0.0137	9537
48	LM	3	0.0336 *	-0.0392 **				-0.2342 **		0.0020	9538

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table B.2: H1: Regression result for general-company tweets data-set

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the general-company data-set. The score column denote which sentiment analysis tool is used to derive sentiment. The agg column denote which aggregation method is used to construct sentiment time-series: 1, 2 and 3 denote count-neutral, count-dependent and mean-sentiment as defined in Equations 4.2, 4.3 and 4.4. Sent is the sentiment. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated with Whites HC robust covariance estimator. The stars denote significance level.

reg #	Score	Agg	Intercept	r_lag_1	rm	rm_lag_1	sent	sent_lag_1	sent_lag_2	R-Squared	# obs
49	Vader	1	0.0166	-0.0462 *	1.0320 ****	0.0456	0.8274 ****	-0.3653 **	0.1238	0.3879	9300
50	Vader	1	0.0163	-0.0438 *	1.0354 ****	0.0438		0.0491	0.1623	0.3842	9300
51	Vader	1	0.0162	-0.0458 *	1.0319 ****	0.0443	0.8362 ****	-0.2984 **		0.3869	9330
52	Vader	1	0.0160	-0.0435 *	1.0354 ****	0.0424		0.1411		0.3831	9330
53	Vader	1	0.0318 *	-0.0450 **			1.1070 ****	-0.6988 ****	0.1745	0.0088	9300
54	Vader	1	0.0315 *	-0.0427 **				-0.1448	0.2265	0.0022	9300
55	Vader	1	0.0336 *	-0.0447 **			1.1186 ****	-0.6017 ****		0.0086	9330
56	Vader	1	0.0334 *	-0.0426 **				-0.0140		0.0018	9330
57	Vader	2	0.0165	-0.0458 *	1.0326 ****	0.0455	0.2723 ****	-0.1251 **	0.0617	0.3869	9300
58	Vader	2	0.0163	-0.0436 *	1.0355 ****	0.0439		0.0065	0.0789 *	0.3843	9300
59	Vader	2	0.0162	-0.0454 *	1.0324 ****	0.0441	0.2754 ****	-0.0887		0.3858	9330
60	Vader	2	0.0160	-0.0433 *	1.0355 ****	0.0422		0.0533		0.3831	9330
61	Vader	2	0.0317 *	-0.0448 **			0.3850 ****	-0.2514 ****	0.0601	0.0074	9300
62	Vader	2	0.0315 *	-0.0427 **				-0.0656	0.0845	0.0022	9300
63	Vader	2	0.0336 *	-0.0444 **			0.3876 ****	-0.2134 ***		0.0072	9330
64	Vader	2	0.0334 *	-0.0425 **				-0.0137		0.0018	9330
65	Vader	3	0.0164	-0.0443 *	1.0324 ****	0.0452	1.5895 ****	-1.0316 **	0.2486	0.3862	9300
66	Vader	3	0.0163	-0.0428 *	1.0351 ****	0.0431		-0.1996	0.3577	0.3840	9300
67	Vader	3	0.0161	-0.0439 *	1.0324 ****	0.0439	1.6107 ****	-0.8446 **		0.3852	9330
68	Vader	3	0.0160	-0.0425 *	1.0351 ****	0.0417		0.0615		0.3830	9330
69	Vader	3	0.0315 *	-0.0432 **			2.3089 ****	-1.9493 ****	0.2697	0.0069	9300
70	Vader	3	0.0315 *	-0.0422 **				-0.7423 *	0.4291	0.0024	9300
71	Vader	3	0.0334 *	-0.0427 **			2.3267 ****	-1.7391 ****		0.0067	9330
72	Vader	3	0.0333 *	-0.0419 **				-0.4310		0.0021	9330
73	LM	1	0.0169	-0.0482 **	1.0323 ****	0.0485 *	0.7682 ****	-0.2356 **	-0.1121	0.3901	9300
74	LM	1	0.0164	-0.0451 *	1.0360 ****	0.0443		0.1416	-0.0351	0.3842	9300
75	LM	1	0.0165	-0.0469 *	1.0320 ****	0.0476 *	0.7655 ****	-0.2887 ***		0.3892	9330
76	LM	1	0.0160	-0.0441 *	1.0358 ****	0.0433		0.1299 *		0.3832	9330
77	LM	1	0.0321 *	-0.0451 **			0.9415 ****	-0.5068 ****	-0.1090	0.0109	9300
78	LM	1	0.0315 *	-0.0433 **				-0.0446	-0.0142	0.0020	9300
79	LM	1	0.0339 *	-0.0437 **			0.9370 ****	-0.5491 ****		0.0109	9330
80	LM	1	0.0334 *	-0.0423 **				-0.0364		0.0018	9330
81	LM	2	0.0169	-0.0473 **	1.0322 ****	0.0478 *	0.3314 ****	-0.1280 **	-0.0321	0.3899	9300
82	LM	2	0.0164	-0.0444 *	1.0358 ****	0.0439		0.0390	-0.0022	0.3840	9300
83	LM	2	0.0165	-0.0463 *	1.0320 ****	0.0469 *	0.3317 ****	-0.1398 ***		0.3891	9330
84	LM	2	0.0160	-0.0437 *	1.0356 ****	0.0429		0.0442		0.3831	9330
85	LM	2	0.0321 *	-0.0444 **			0.4033 ****	-0.2478 ****	-0.0222	0.0108	9300
86	LM	2	0.0315 *	-0.0427 **				-0.0446	0.0143	0.0020	9300
87	LM	2	0.0340 *	-0.0433 **			0.4034 ****	-0.2501 ****		0.0107	9330
88	LM	2	0.0334 *	-0.0421 **				-0.0262		0.0019	9330
89	LM	3	0.0168	-0.0457 *	1.0332 ****	0.0467 *	1.7434 ****	-0.8826 ***	-0.0459	0.3892	9300
90	LM	3	0.0163	-0.0434 *	1.0353 ****	0.0432		0.0334	0.0902	0.3840	9300
91	LM	3	0.0165	-0.0450 *	1.0331 ****	0.0459 *	1.7704 ****	-0.8769 ***		0.3885	9330
92	LM	3	0.0160	-0.0430 *	1.0352 ****	0.0423		0.1321		0.3830	9330
93	LM	3	0.0320 *	-0.0435 **			1.9908 ****	-1.3798 ****	0.0403	0.0089	9300
94	LM	3	0.0315 *	-0.0425 **				-0.3325	0.1956	0.0021	9300
95	LM	3	0.0340 *	-0.0428 **			2.0165 ****	-1.3001 ****		0.0090	9330
96	LM	3	0.0334 *	-0.0422 **				-0.1497		0.0019	9330

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table B.3: H2: Regression result for combined data-set, Vader

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the both data-sets. The score column denote which sentiment analysis tool is used to derive sentiment. The agg column denote which aggregation method is used to construct sentiment time-series: 1, 2 and 3 denote count-neutral, count-dependent and mean-sentiment as defined in Equations 4.2, 4.3 and 4.4. Gc_Sent is the sentiment from the general-company data-set. Cash_sent is the sentiment from the cashtags data-set. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated with Whites HC robust covariance estimator. The stars denote significance level.

reg#	Scoring	Agg	Intercept	r_lag_1	rm	rm_lag_1	gc_sent	gc_sent_lag_1	gc_sent_lag_2	cash_sent	cash_sent_lag_1	cash_sent_lag_2	R-squared	# obs
97	Vader	1	0.0172	-0.0534 **	1.0289 ****	0.0482 *	0.7114 ****	-0.3403 **	0.1286	0.3253 ****	-0.0449	-0.0301	0.3914	9300
98	Vader	1	0.0164	-0.0440 *	1.0356 ****	0.0440		0.0460	0.1644		0.0091	-0.0026	0.3843	9300
99	Vader	1	0.0167	-0.0529 **	1.0289 ****	0.0471 *	0.7189 ****	-0.2756 **		0.3265 ****	-0.0501		0.3904	9330
100	Vader	1	0.0161	-0.0436 *	1.0356 ****	0.0425		0.1385			0.0090		0.3832	9330
101	Vader	1	0.0324 *	-0.0524 ***			0.9552 ****	-0.6305 ****	0.2024	0.4347 ****	-0.1510 ***	-0.1032 **	0.0154	9300
102	Vader	1	0.0315 *	-0.0420 **				-0.1128	0.2515 *		-0.0792	-0.0667	0.0026	9300
103	Vader	1	0.0342 *	-0.0515 ***			0.9655 ****	-0.5368 ***		0.4312 ****	-0.1643 ****		0.0150	9330
104	Vader	1	0.0334 *	-0.0414 **				0.0185			-0.0866 *		0.0021	9330
105	Vader	2	0.0175	-0.0553 **	1.0289 ****	0.0477 *	0.2158 ****	-0.1078 *	0.0654	0.2042 ****	-0.0398 **	-0.0294	0.3927	9300
106	Vader	2	0.0164	-0.0438 *	1.0356 ****	0.0439		0.0058	0.0813 *		0.0034	-0.0067	0.3843	9300
107	Vader	2	0.0170	-0.0545 **	1.0291 ****	0.0465 *	0.2180 ****	-0.0727		0.2031 ****	-0.0468 ***		0.3916	9330
108	Vader	2	0.0161	-0.0433 *	1.0356 ****	0.0423		0.0531			0.0016		0.3831	9330
109	Vader	2	0.0328 *	-0.0545 ***			0.3175 ****	-0.2177 ***	0.0729	0.2503 ****	-0.0893 ****	-0.0675 ***	0.0166	9300
110	Vader	2	0.0316 *	-0.0419 **				-0.0533	0.0959 *		-0.0360	-0.0395	0.0027	9300
111	Vader	2	0.0345 *	-0.0529 ***			0.3193 ****	-0.1815 ***		0.2456 ****	-0.1014 ****		0.0159	9330
112	Vader	2	0.0334 *	-0.0411 **				-0.0011			-0.0425 *		0.0021	9330
113	Vader	3	0.0174	-0.0528 **	1.0287 ****	0.0471 *	1.2907 ****	-0.9315 **	0.2928	1.5494 ****	-0.2143	-0.2303	0.3910	9300
114	Vader	3	0.0166	-0.0431 *	1.0348 ****	0.0434		-0.2107	0.3772		0.0543	-0.0933	0.3838	9300
115	Vader	3	0.0168	-0.0518 **	1.0291 ****	0.0459	1.3071 ****	-0.7285 *		1.5430 ****	-0.2784 **		0.3902	9330
116	Vader	3	0.0162	-0.0425 *	1.0352 ****	0.0419		0.0597			0.0146		0.3830	9330
117	Vader	3	0.0327 *	-0.0520 ***			1.9387 ****	-1.7624 ****	0.3395	1.9328 ****	-0.5831 ***	-0.4286 **	0.0151	9300
118	Vader	3	0.0320 *	-0.0415 **				-0.6954	0.4691		-0.2497	-0.2599	0.0027	9300
119	Vader	3	0.0340 *	-0.0506 ***			1.9470 ****	-1.5337 ****		1.9111 ****	-0.6677 ****		0.0146	9330
120	Vader	3	0.0334 *	-0.0407 **				-0.3760			-0.3069 *		0.0023	9330

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table B.4: H2: Regression result for combined data-set, LM

In this table we present the result of the regressions that try to model the relationship between company-level log returns and sentiment derived from the both data-sets. The score column denote which sentiment analysis tool is used to derive sentiment. The agg column denote which aggregation method is used to construct sentiment time-series: 1, 2 and 3 denote count-neutral, count-dependent and mean-sentiment as defined in Equations 4.2, 4.3 and 4.4. Gc_Sent is the sentiment from the general-company data-set. Cash_sent is the sentiment from the cashtags data-set. rm is the S&P500 index log return. r is the company-level log return. The parameters are estimated using panel regression with fixed entity effects on daily observations. P-values are calculated with Whites HC robust covariance estimator. The stars denote significance level

#	Scoring	Agg	Intercept	r_lag_1	rm	rm_lag_1	gc_sent	gc_sent_lag_1	gc_sent_lag_2	cash_sent	cash_sent_lag_1	cash_sent_lag_2	R-squared	# obs
121	LM	1	0.0182	-0.0558 **	1.0338 ****	0.0494 *	0.7671 ****	-0.2392 **	-0.1423	0.2470 ****	-0.0734	-0.0235	0.3923	9300
122	LM	1	0.0139	-0.0462 *	1.0401 ****	0.0440		0.1468	-0.0408		-0.0253	-0.0031	0.3828	9300
123	LM	1	0.0189	-0.0526 **	1.0338 ****	0.0484 *	0.7375 ****	-0.2914 **		0.2401 ****	-0.0782		0.3915	9330
124	LM	1	0.0142	-0.0441 *	1.0393 ****	0.0434		0.1303			-0.0277		0.3822	9330
125	LM	1	0.0322 *	-0.0536 ***			0.9202 ****	-0.4883 ****	-0.1031	0.3274 ****	-0.1525 ***	-0.0939 **	0.0211	9300
126	LM	1	0.0287 *	-0.0442 **				-0.0128	0.0164 *		-0.0914	-0.0618	0.0032	9300
127	LM	1	0.0354 *	-0.0490 ***			0.8862 ****	-0.5111 ***		0.3085 ****	-0.1716 ****		0.0197	9330
128	LM	1	0.0313 *	-0.0411 **				0.0083			-0.1068 *		0.0028	9330
129	LM	2	0.0181	-0.0586 **	1.0297 ****	0.0493 *	0.2987 ****	-0.1196 *	-0.0414	0.2105 ****	-0.0581 **	-0.0270	0.3955	9300
130	LM	2	0.0138	-0.0454 *	1.0398 ****	0.0435		0.0385	-0.0031 *		-0.0128	-0.0050	0.3827	9300
131	LM	2	0.0189	-0.0554 **	1.0302 ****	0.0485 *	0.2880 ****	-0.1280		0.2067 ****	-0.0650 ***		0.3947	9330
132	LM	2	0.0142	-0.0435 *	1.0392 ****	0.0430		0.0438			-0.0148		0.3821	9330
133	LM	2	0.0320 *	-0.0573 ***			0.3493 ****	-0.2214 ***	-0.0127	0.2864 ****	-0.1165 ****	-0.0770 ***	0.0278	9300
134	LM	2	0.0286 *	-0.0429 **				-0.0308	0.0322 *		-0.0574	-0.0447	0.0035	9300
135	LM	2	0.0353 *	-0.0522 ***			0.3378 ****	-0.2117 ***		0.2742 ****	-0.1350 ****		0.0261	9330
136	LM	2	0.0313 *	-0.0399 **				-0.0034			-0.0699 *		0.0029	9330
137	LM	3	0.0177	-0.0538 **	1.0284 ****	0.0499 *	1.4511 ****	-0.7680 **	-0.0631	0.9573 ****	-0.1608	-0.1371	0.3942	9300
138	LM	3	0.0166	-0.0438 *	1.0351 ****	0.0435		0.0203	0.1018		0.0420	-0.0308	0.3837	9300
139	LM	3	0.0170	-0.0523 **	1.0290 ****	0.0491	1.4793 ****	-0.7876 *		0.9385 ****	-0.1916 **		0.3936	9330
140	LM	3	0.0161	-0.0432 *	1.0354 ****	0.0426		0.1231			0.0312		0.3831	9330
141	LM	3	0.0330 *	-0.0516 ***			1.6118 ****	-1.1533 ****	0.0594	1.2596 ****	-0.4473 ***	-0.3454 **	0.0186	9300
142	LM	3	0.0320 *	-0.0417 **				-0.2708	0.2527		-0.1847	-0.2052	0.0027	9300
143	LM	3	0.0342 *	-0.0492 ***			1.6455 ****	-1.1065 ****		1.2180 ****	-0.5247 ****		0.0180	9330
144	LM	3	0.0334 *	-0.0405 **				-0.0795			-0.2392 *		0.0022	9330

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table B.5: H3: Volatility regression result for cashtags tweets data-set

In this table we present the result of the regressions that try to model the relationship between company-level conditional volatility and select sentiment-features derived from the cashtags data-set. Sent^2 is the squared count-dependent LM sentiment. #total is the total number of tweets. #pos and #neg are the number of positive and negative tweets. Cond_vol is the conditional volatility as estimated by a constant mean GARCH(1,1)-model. resid are the residuals from the same estimation. The parameters are estimated using panel regression with fixed entity effects on daily observations (note that we do not impose restrictions of parameter signs in these regressions as the conditional volatility is already estimated. P-values are calculated using Driscol Kraay HAC robust covariance estimator. The stars denote significance level.

reg #	Score	Agg	Intercept	condvol_1_lag_1	resid_lag_1	sent^2_lag_1	#total_lag_1	#pos_lag_1	#neg_lag_1	sent_var_lag_1	R-squared	# obs
145	Vader	1	0.0116	0.8011 ****	-0.4321 **	0.0019	0.0075 ****	0.0020	0.0030 ***	0.0158	0.6867	9294
146	Vader	1	0.0480 ****	0.8166 ****	-0.4406 **	-0.0047 ***					0.6774	9330
147	Vader	2	0.0114	0.8012 ****	-0.4330 **	-0.0006	0.0086 ****	0.0010	0.0027 ***	0.0154	0.6867	9294
148	Vader	2	0.0480 ****	0.8168 ****	-0.4398 **	-0.0013 ***					0.6774	9330
149	Vader	3	0.0117	0.8011 ****	-0.4322 **	0.0067	0.0080 ***	0.0015	0.0030 ***	0.0150	0.6866	9294
150	Vader	3	0.0478 ****	0.8168 ****	-0.4398 **	-0.0526 **					0.6770	9327
151	LM	1	0.0475 ****	0.7998 ****	-0.4365 **	0.0011	-0.0052	0.0110 ****	0.0102 ****	-0.0526 ****	0.6871	9078
152	LM	1	0.0486	0.8167 ****	-0.4368 **	-0.0035 ***					0.6762	9108
153	LM	2	0.0480 ****	0.7999 ****	-0.4373 **	-0.0001	-0.0051	0.0110 ****	0.0099 ****	-0.0533 ****	0.6871	9078
154	LM	2	0.0479	0.8169 ****	-0.4373 **	-0.0004					0.6760	9108
155	LM	3	0.0480 ****	0.7999 ****	-0.4373 **	-0.0001	-0.0051	0.0110 ****	0.0099 ****	-0.0533 ****	0.6871	9078
156	LM	3	0.0479 ****	0.8169 ****	-0.4373 **	-0.0004					0.6760	9108
157			0.0083	0.8029 ****	-0.4389 **		0.0119 ****				0.6864	9300
158	Vader		0.0237 ****	0.8071 ****	-0.4491 **			0.0097 ****			0.6848	9300
159	Vader		0.0339 ****	0.8034 ****	-0.4158 **				0.0091 ****		0.6844	9300
160	Vader		0.0430 ****	0.8156 ****	-0.4374 **					0.0473 ****	0.6774	9324
161	LM		0.0356 ****	0.8077 ****	-0.4611 **			0.0094 ****			0.6844	9300
162	LM		0.0349 ****	0.8061 ****	-0.4177 **				0.0082 ****		0.6832	9300
163	LM		0.0466 ****	0.8168 ****	-0.4386 **					0.0034	0.6769	9324

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001

Table B.6: H3: Volatility regression result for general-company tweets data-set

In this table we present the result of the regressions that try to model the relationship between company-level conditional volatility and select sentiment-features derived from the general-company data-set. Sent^2 is the squared count-dependent LM sentiment. #total is the total number of tweets. #pos and #neg are the number of positive and negative tweets. Cond_vol is the conditional volatility as estimated by a constant mean GARCH(1,1)-model. resid are the residuals from the same estimation. The parameters are estimated using panel regression with fixed entity effects on daily observations (note that we do not impose restrictions of parameter signs in these regressions as the conditional volatility is already estimated. P-values are calculated using Driscol Kraay HAC robust covariance estimator. The stars denote significance level.

reg #	Score	Agg	Intercept	condvol_1_lag_1	resid_lag_1	sent^2_lag_1	#total_lag_1	#pos_lag_1	#neg_lag_1	sent_var_lag_1	R-squared	# obs
164	Vader	1	0.0308 *	0.8097 ****	-0.4333 **	0.0254 *	0.0027	-0.0012	0.0019	-0.0370	0.6703	9300
165	Vader	1	0.0479 ****	0.8100 ****	-0.4328 **	0.0353 **					0.6695	9330
166	Vader	2	0.0343 **	0.8092 ****	-0.4314 **	0.0045 **	0.0024	-0.0023	0.0029	-0.0383	0.6705	9300
167	Vader	2	0.0478 ****	0.8098 ****	-0.4323 **	0.0059 ***					0.6696	9330
168	Vader	3	0.0358 **	0.8093 ****	-0.4327 **	0.1086 *	0.0019	-0.0022	0.0036	-0.0512 *	0.6703	9300
169	Vader	3	0.0481 ****	0.8101 ****	-0.4349 **	0.1381 **					0.6692	9330
170	LM	1	0.0189	0.8096 ****	-0.4332 **	0.0210 **	0.0045	-0.0012	-0.0001	0.0051	0.6704	9300
171	LM	1	0.0475 ****	0.8100 ****	-0.4330 **	0.0257 ***					0.6696	9330
172	LM	2	0.0175	0.8093 ****	-0.4317 **	0.0034 ***	0.0054	-0.0017	-0.0007	0.0102	0.6704	9300
173	LM	2	0.0478 ****	0.8097 ****	-0.4313 **	0.0041 ****					0.6698	9330
174	LM	3	0.0106	0.8093 ****	-0.4329 **	0.0687 **	0.0075	-0.0024	-0.0017	0.0138	0.6703	9300
175	LM	3	0.0481 ****	0.8097 ****	-0.4325 **	0.0844 ***					0.6695	9330
176			0.0135	0.8105 ****	-0.4380 **		0.0046 **				0.6698	9300
177	Vader		0.0235 *	0.8111 ****	-0.4400 **			0.0036 **			0.6695	9300
178	Vader		0.0266 ***	0.8098 ****	-0.4351 **				0.0037 ***		0.6698	9300
179	Vader		0.0516 ****	0.8114 ****	-0.4385 **					-0.0193	0.6687	9330
180	LM		0.0285 **	0.8108 ****	-0.4415 **			0.0035 *			0.6696	9300
181	LM		0.0296 ***	0.8102 ****	-0.4350 **				0.0032 **		0.6696	9300
182	LM		0.0460 ****	0.8110 ****	-0.4377 **					0.0078	0.6687	9330

* indicate significance of 0.05 to 0.1
** indicate significance of 0.01 to 0.05
*** indicate significance of 0.001 to 0.01
**** indicate significance of < 0.001