Master Thesis in Finance - 4350

Agne Macijauskaite - 40992
Laurynas Ruzgas - 40991

Michael Halling - Supervisor
2018-05-14

Twitter based sentiment effect on stock market returns

- sentiment analysis of Twitter data and its effect on the U.S. stock market: perspective of S&P 500, industries and investor types

---

ABSTRACT. The purpose of this study is to investigate the effect of Twitter based sentiment on stock market returns. We use a uniquely large dataset of 95 million tweets concerning the 102 biggest US companies, S&P 500 index and other market keywords for the year 2017. Correspondingly, the dataset also includes the returns from our sample companies and the S&P 500 index. The methodology used in this study comprises of several steps, but above all, 1) we use two dictionary based sentiment analysis tools for converting the tweets to sentiment scores, and 2) we use an autoregressive distributed lag (ADL) model for the empirical analysis. Our contribution to the literature can be summarized with the following four key findings: a) the newer data confirmed that Twitter sentiment, especially the bullishness index, has predictive power of S&P 500 returns; b) Vader NLTK sentiment analysis outperforms the traditionally used Loughran and McDonald Lexicon-based method; c) some industry returns have higher sensitivity to sentiment, while predictive power was only found for the IT industry, up to 2 days ahead; and d) the conventional opinion that retail investors are more affected by sentiment is not confirmed, on the contrary, we find that sentiment has more predictive power for companies with a high institutional investor share.

Key words: sentiment, Twitter, S&P 500, returns, stock market, industries, investor type

***Acknowledgements**: While remaining responsible for any errors in this thesis, the authors are thankful to Michael Halling for his excellent support and valuable advice.

Contents

1 Introduction

The knowledge of which factors can explain and predict asset prices is highly valuable for investors, regulators and other parties, thus it is not a coincidence that this research field is well studied in academia. Consequently, well known asset pricing models have emerged over time such as the Efficient Market Hypothesis (EMH), developed by Nobel prize winner Eugene Fama (1970), that stated that asset prices reflect all publicly available information and should react only to new relevant information. Thus, artifacts as feelings, emotions or opinions, i.e. sentiment, should not have any effect on asset prices.

However, more than half a century ago Stone, Dunphy, Smith, and Ogilvie (1968) described how texts with emotional information, or sentiment, could be essential for the prediction of investor behavior and thus asset prices. Despite this, not until the 1980s, serious attempts were made to explore the possibility that the market was not as efficient as the theory predicted. Within behavioral finance, theories started to develop that explained how sentiment might have an effect on asset prices. Since then, a growing body of empirical research has emerged, trying to explore the relationship between sentiment and asset prices and a lot of results have been found. (Brown & Cliff, 2004)

To begin with, research investigated traditional sources for sentiment such as company disclosures and news-media, which were the only mediums at the time. With the rise of the Internet, new mediums arose as blogs, message boards and social media, i.e. Twitter. Twitter is the largest micro-blogging platform in the world and has received more research attention after the highly significant results were published by Bollen, Mao, and Zeng (2011). They found that certain types of sentiment had significant predictive power of DJIA returns for up to five days ahead. They also expanded the model including a neural networks and found an 87% prediction accuracy of the returns direction. Since then, the study has become an inspiration for a lot of research to come.

In this study we continue to explore the connection between Twitter based sentiment and stock market returns. Specifically, we investigate whether 1) previous found results can be replicated with new data as well as using different methodologies, 2) certain industries can be better predicted than others, and 3) there is a difference in predictability among companies with large share of retail-to-institutional investors. These areas are investigated using a traditional and extended methodology with data for 2017 comprising of around 95 million top tweets.

The remainder of this paper is composed as follows. Section 2 provides an extended literature review explaining the development of the sentiment related research in connection with asset prices, its main empirical findings and ends with our hypotheses. Section 3 presents our Twitter- and Financial data. Section 4 explains methodology applied in this study and the extensions we made. Section 5 presents our results and discussions and lastly, Section 6 outlines some concluding remarks and proposals for further research.

2 Literature overview and our contribution

A lot of research have investigated whether new relevant information affects asset prices as the efficient market hypothesis predicts and plenty of supporting results have been found over the years. In connection with EMH, researchers took different perspectives: some tried to prove the efficiency of the market and the degree of it, while the others took EMH as a fact and tried to identify which news are relevant in predicting asset returns. However, the backbone of this research says that irrelevant information for asset fundamentals should not affect its prices (Fama, 1970). Thus, the emotions embedded in newly published news or the feeling towards the aggregated market should not have any price effects (Brown & Cliff, 2004). However, more than half a century ago Stone et al. (1968) described how texts with emotional information, or sentiment, could be essential for the prediction of investor behavior and thus asset prices. But what is sentiment and how could it be relevant for asset prices?

In this paper we describe sentiment as a feeling, attitude, emotion or opinion of an individual person or a group (i.e. aggregated market). The sentiment can be expressed in many ways, while in this paper we focus only on sentiment expressed in texts. The connection between sentiment and asset prices is explored through the field of behavioral finance and this relationship started to develop as late as in the mid 1980s, at a time when serious attempts were made to explore the possibility that liquid financial markets were not as stable as was predicted by the efficient market theory (Brown & Cliff, 2004). In order to describe further developments of this field, we structure the Literature overview part in several sections. We start by outlying the theories that attempt to explain why the relationship between sentiment and asset prices should exist or exists. Further, we focuses on the existing financial literature, which explored the relationship between asset prices and sentiment from different mediums. Lastly, we provide a more focused literature overview that investigated the relationship between asset returns and our chosen sentiment medium - Twitter. We conclude this section by focusing on our research questions and their motivation.

2.1 Reasons for sentiment effect on asset prices

As mentioned, the Efficient Market Hypothesis assumes that all relevant information is incorporated in asset prices, and thus the price movements are unpredictable. However, many studies found market inefficiencies. One of the common theory that talks about prolonged inefficiencies for asset prices is the noise trading theory. It explains why mis-pricing of assets can persist over a period of time and only after a while the prices rebound to its fundamentals (De Long, Shleifer, Summers, & Waldmann, 1990; Shleifer & Vishny, 1997). The most common underlying reason for prolonged mis-pricing is 'noise trader risk', which implies that arbitrageurs do not correct the market straight away because they can lose money in the short term. This risk is especially pronounced in small companies, as relevant information for large ones gets incorporated significantly faster (Hong & Stein, 1999). It can be concluded that the noise trading theory explains asset mis-pricing over a period of time, however, it does not say how people actually form their beliefs and expectations for it to happen in the first place.

There is no consistent explanation how people form their beliefs in the literature, however, it is possible that sentiment might relate to several existent interpretations. In relation to this, Barberis, Shleifer, and Vishny (1998) use well known behavioral biases, to be more precise: representation and conservatism, to explain how people form their beliefs and how it leads to asset mis-pricing over a period of time. Representation bias refers to people forming beliefs about new information and expressed opinions mainly in relation to their prior beliefs about the asset or asset class and neglects the actual statistical probabilities. For example, new spreading popular opinions, which are in line with the investors' initial assumptions about that asset, could lead to an overreaction. While conservatism relates to slow updating of publics' expectations in relation to new information. In the financial-markets conservatism leads to investors updating their strategies in the right direction, just in lower magnitude, which in turn leads to underreaction. Griffin and Tversky (1992) combined these two biases and proposed a theory that people base their decisions on 'strength' and 'weight', where 'strength' refers to emotional importance and salience of the news, while 'weight' refers to the statistical properties of the event. According to Griffin and Tversky (1992), people put too much importance on the 'strength' and care substantially too little about the 'weight'. Thus, underreaction would occur when news are not that salient, but fundamentally important (high 'strength'), which is in line with the conservatism bias. Contrary, the 'overreaction' would occur when news or opinions are popular and spreading fast, while the real relevance to fundamental prices are low, which is in line with representation bias. Both of these biases would be expressed as well as exploited by noise traders and would be corrected by arbitrageurs only after some periods. The research of these biases, especially through sentiment, is becoming more and more common. We are going to discuss the existent literature in the subsequent section.

2.2 Asset prices and sentiment from different mediums

From an empirical point of view, a lot of research have explored the connection between sentiment and asset prices. Kearney and Liu (2014) have done an extensive summary within this field and found that sentiment sources used in the literature are predominately from: public corporate disclosures/filings, media articles and Internet messages, where the later source is the least studied but rapidly growing. All mediums come with different advantages, disadvantages and characteristics.

The characteristics of corporate disclosures are that they come embedded with relevant fundamental information together with sentiment. This presence of fundamental information within the medium exposes it to both, underreaction and overreaction possibilities discussed in the previous section. In addition, the disadvantage of corporate disclosures medium is low reporting frequency and thus limited data. Due to this limitation, most researchers conducted event studies in relation to different types of corporate announcements. The main conclusions were that mood changes in these disclosures compared to previous filings had a significant contemporaneous effect on returns, even after controlling for surprises in fundamentals. Thus, even controlling for relevant information, this source of sentiment had an effect on asset prices. (Kearney & Liu, 2014)

Media-expressed sentiment refers to the news articles with relevant information. This medium focuses on reporting information rather than expressing new opinions, therefore it could be said that information provided is more related to the past or current events, rather than future prospects (Kearney & Liu, 2014). One of the main contributors within this sentiment data source is Tetlock (2007). In the paper he found that high media pessimism predicts negative pressure on market index prices. The effect is followed by reversion to fundamentals over a few days. While for small stocks, the price impact of negative sentiment is even larger and reversal is slower. He concluded that the found results were consistent with the theoretical noise trading models and that the sampled media information was not providing new relevant fundamental information.

With the rise of the Internet, a completely new medium - social media - was born with very different characteristics. Here we find blogs, message boards and micro-blogs, including Twitter. We will discuss literature concerning Twitter in section 2.3. However, talking about social media data, it is important to note that it is far more abundant but simultaneously contains substantially more noise with respect to new relevant information. This makes it an interesting source to show market inefficiencies and confirm some behavioral finance theories (Kearney & Liu, 2014)[1]. The lack of fundamental information within the social media platforms and the ease of spreading (high salience) expose this medium significantly to investors' overreaction and less so to underreaction (Barberis et al., 1998). Additionally, Yu, Duan, and Cao (2013) found that social-media is a better predictor of stock returns than conventional media, when only regarding sentiment. Thus, social media is 1) the newest of the three discussed data sources, 2) the most relevant for behavioral sentiment analysis, especially in relation to overreaction, 3) as well as it has been found to be the best for predicting the stock market, with regards only to the sentiment. Therefore, we have chosen to focus our paper on social media source. Antweiler and Frank (2004) is one of the first well-cited papers within this medium. They found that positive shocks in message boards postings predict negative returns for the next day. The effect on stock returns was statistically significant but economically small. However, more interestingly this study also used intraday data (15 min intervals) in modeling sentiment and stock market movements, which is unique as daily data is by far the most common time frame used in the literature. They found statistically significant predictive power of sentiment even in 15 min intervals. Following these findings, and the rise of new social platforms, research using these mediums have expanded significantly.

The rise of popularity in social media sources increased the importance of deciding which source of data is the most effective within the group. H. Mao, Counts, and Bollen (2011) compared sentiment from different social media as well as conventional media sources and their predictive power on market returns. They found that Google search volume of financial terms as well as Twitter (both sentiment and volume) were the most significant predictors of returns. They also found that Twitter volumes were earlier indicator of big changes compared to Google search volumes. This was again confirmed in their later study,

[1] It could be argued that texts in social media express lagging reaction to fundamental information. Even if that is true, their predictive power would indicate slow information incorporation in asset prices - thus low efficiency of the market.

which found that Google search volumes lag Twitter sentiment (H. Mao, Counts, & Bollen, 2015). Following these results, we decided to focus on Twitter as a data source for extracting sentiment in our paper.

2.3 Twitter and asset prices

In regards to Twitter (www.twitter.com), it is the largest micro-blogging platform in the world. It allows the users to share the tweets, which are short messages of up to 140 characters (recently changed to 280). The information posted is visible publicly through their website. People can search for specific messages using the search query of specific words or terms. This micro-blogging platform grew significantly over the years and now have a total amount of 650 mn registered users and 500 mn newly generated tweets a day (Aslam, 2018). The tweet text might involve both business related information as well as general financial market-irrelevant statements. Many institutional and retail investors, as well as analysts, post news and investment opinions on Twitter, providing a more extensive news media than the traditional outlets (Sprenger, Tumasjan, Sandner, & Welpe, 2014a). Due to its growth in general public popularity, the research using the medium expanded significantly.

The first well-cited study to test the connection between Twitter sentiment and stock market returns was conducted by Bollen et al. (2011). They collected the tweets about the general public feeling using search queries, such as 'I am feeling', and investigated different text sentiment dimensions. They found that certain types of sentiment, specifically a measure they called 'calm', had significant predictive power for DJIA returns up to five days ahead. They also expanded the model including a neural networks and found an 87% prediction accuracy of returns direction, thus concluding that the connection between twitter sentiment and returns may not be linear. After these results, this study has become an inspiration for many of the studies going forward.

However, as this medium is fairly new, the relationship between Twitter sentiment and the stock market is still in its early days, but rapidly growing. Additionally, due to technical improvements in sentiment extractions from text, the research field that includes text analysis or sentiment analysis is also growing. In Table 1, we made a comprehensive summary of the main studies that researched the relationship between Twitter sentiment and the stock market - our paper's focus area. We discuss these findings in more details in the following part.

Table 1: Literature review

The table includes the literature that covers the relationship between Twitter data and asset returns. The Research column includes the author and the date of the publication. The Data column shows the number of data points (if provided in the paper), the used medium for sentiment and the used financial data. The Time frame includes the time period used for testing the data, both the year and number of months, as well as data frequency. In the Methodology section, one can found which sentiment analysis as well as which empirical analysis were used in modeling the relationship between dependent and independent variables. In Results section we provide the summary of the key results of the papers.

#	Research	Data	Time frame	Methodology	Results
1	Bollen et al. (2011)	9.9 mn messages of mood; Twitter; DJIA	2008 (10M): Daily	Dictionary-based; Linear regression + SOFNN	Changes in the public mood can be tracked using Twitter data. Some mood dimensions, especially calm and happy, have a predictive power of DJIA returns. Changes of some mood dimensions are significant predictors of the DJIA between 1-5 days in advance. Public mood has a non linear relationship with DJIA.
2	Zhang X. et al. (2011)	5.5 mn messages of mood; Twitter; DJIA, S&P 500, Nasdaq	2009 (5M): Daily	No sentiment analysis; Correlation	Mood words predict market returns. Words such as Hope, Worry, Fear and Anxious are the most significantly correlated words with next day market returns. Combination of negative words has stronger predictive power than positive cobinations. Number of emotional words (both negative and positive) has negative correlation with next day returns and positive correlation with volatility.
3	Mao et al. (2011)	Twitter, Google search, Surveys, News; DJIA, VIX, gold	2010-2011 (15M): Daily; 2008-2011 (33M): Weekly	No sentiment analysis; Linear regression	Sentiment from all mediums were correlated with returns and VIX on a daily basis. Twitter bullishness and the volume of tweets are significant predictors of returns (1 and 2 lags). Survey data was not significant and News data had lower significance. Google search volumes were predictive on a weekly basis. It also found that Twitter volumes precede Google search volumes.
4	Mittal & Goel (2012)	Messages of mood; Twitter; DJIA	2009 (7M): Daily	Dictionary-based; Linear regression + non-linear transformations	Public mood can be captured using large scale Twitter data. Only the moods 'calm' and 'happy' has a causal relationship with DJIA at 3-4 lags. SOFNN outperforms other techniques.
5	Ruiz et al. (2012)	Twitter; 150 companies of S&P 500	2010 (6M): Daily	Dictionary-based; Time-constrained graphs & correlation	The most correlated components are the number of connected components and the amount of nodes in the interaction graphs. Sentiment has higher correlation with trading volume rather than returns. Even with this small correlation to the returns, a profitable trading strategy was posible.
6	Si et al. (2013)	0.62 mn messages of companies; Twitter; S&P 100	2012-2013 (3M): Daily	Dictionary-based; VAR	Topic-based public sentiment can improve the accuracy of stock prediction compared to non topic-based sentiment approaches. This implies that some topics have a higher impact on stock returns.
7	Smajlović et al. (2013)	0.15 mn messages of companies; Twitter; 8 companies	2011 (9M): Daily	Machine Learning; Linear regression	Changes in positive sentiment probability can predict stock returns, especially in cases of high stock price variations or significant fall in returns. Neutral tweet sentiment in some cases provided additional information in modelling stock returns.
8	Chen & Lazer (2013)	Twitter; Market returns	Daily	Dictionary-based; Linear regression	Twitter sentiment is correlated with the market. Several trading strategies incorporating sentiment data provide profitable results.
9	Yu et al. (2013)	0.05 mn messages of companies; Twitter, blogs, forums and media; 824 companies	2011 (3M): Daily	Dictionary-based; Panel regression	Social media sentiment has a stronger effect on firm stock returns compared to conventional media. Social and conventional media have a significant impact on stock returns. Blog sentiment has a positive coefficient, while forums - negative on modeling returns. Blog and Twitter sentiment has a positive effect on risk.
10	Sprenger et al. (2014a)	0.25 mn messages of companies; Twitter; S&P 100	2010(6M): Daily	Machine Learning; Panel regression	There is a significant association between Twitter data bullishness and returns. It fails to find a significant relationship between bullishness and abnormal returns. Disagreement is associated with increase in trading volume. Tweets with higher quality information were not retweeted more frequently.
11	Sprenger et al. (2014b)	0.44 mn messages of companies; Twitter; S&P 500	2010(6M): Daily	Machine Learning; Event study	Using sentiment data, the study manages to distinguish between bad and good events. It finds that events of good news are more pronounced in comparison to bad news. Also, different type of news events have different significance on returns.
12	Mao et al. (2015)	0.31 mn messages of bullishness; Twitter and Google search; DJIA, S&P 500, Russell 1000; Russell 2000 and other	2010-2012 (36M): Daily	No sentiment analysis; Linear regression	Twitter information precedes Google search queries and is a more powerful predictor of stock market sentiment. Twitter and Google bullishness are positively correlated to investor sentiment in existent surveys. Twitter bullishness predicts positive index returns in the US, UK and Canada. The index price returns to fundamentals within a week.

In the literature overview in Table 1, all the studies use Twitter sentiment to see its relation to stock market metrics. The DJIA and S&P 500 are the two most commonly used financial market indices and a smaller sample of studies investigates individual stocks. Additionally, as the number of tweets reported in these studies (not all studies reported this metric) does not exceed 10 million, compared to our sample of 95 million tweets (see Table 7), we may have the largest dataset to our knowledge. Most studies investigate a timespan of less than one year, thus our 12 months sample is in line with literature. Concerning methodology, these studies do not seem to have any preferred method when it comes to sentiment analysis: machine learning, dictionary based analysis or just counting tweets containing sentiment predefined words are all popular. After tweets have been converted to sentiment, most studies use linear regressions to find the relation with the stock market metrics. VAR models, panel regressions, neural networks and trading strategies are also used while event studies are not as common compared to traditional sentiment sources.

Looking at the main results we see that sentiment from Twitter has a statistical association with the stock market (Bollen et al., 2011; X. Zhang, Fuehres, & Gloor, 2011; H. Mao et al., 2011; Mittal & Goel, 2012; Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012; Si et al., 2013; Smailović, Grčar, Lavrač, & Žnidaršič, 2013; Chen & Lazer, 2013; Yu et al., 2013; Sprenger et al., 2014a; Sprenger, Tumasjan, Sandner, & Welpe, 2014b; H. Mao et al., 2015). Most studies have found that sentiment has a strong same day correlation with market returns, even statistically significant predictive effect has been found by numerous studies. Many studies have found significant sentiment predictive power of 1 day lag or 2 lags and some studies even found predictive power of up to 5 lags (Bollen et al., 2011). However, 1 significant lag is the most common finding. Interestingly, H. Mao et al. (2015) found that sentiment had significant predictive power, however, returns reversed to fundamental values within a week. This is consistent with Tetlock (2007) who found similar results and concluded that it is in line with behavioral noise trading theory. Besides market returns, the small sample that focus on individual stocks also found significant predictability (Ruiz et al., 2012; Smailović et al., 2013; Yu et al., 2013). However, it is important to mention, that the presented studies in Table 1 do not discuss in large extent, whether the sentiment-return relationship is economically significant. Nevertheless, several studies (Bollen et al., 2011; Ruiz et al., 2012; Chen & Lazer, 2013) have conducted forecasting and trading strategies that led to higher portfolio returns. This discussion is a strong indication of sentiment significance in predicting stock returns.

In addition to the significant relationship with returns, there were other important findings in these studies. It seems that tweets containing emotional words showed larger influence on returns. For example, Bollen et al. (2011) who used tweets only containing mood expressions, such as 'I feel', 'I am feeling', had strong results, as well as X. Zhang et al. (2011), who used only emotional words, such as 'anxious', 'worry' and 'hope', had also convincing results. The later study also found that combinations of several words, especially negative ones, had a stronger effect on returns. Similarly, predictability can be improved by focusing on specific topics (Si et al., 2013). H. Mao et al. (2015) expanded the predefined emotion and included the search queries of 'bearish' and 'bullish' words. These terms have already expressed emotions and they have more finance related nature.

In general, the overall results are in line with traditional mediums. However, the common relationship in traditional media sources, which says that negative sentiment has a larger impact than positive, does not seem to be an equally discussed finding within the Twitter medium. Some contradicting results were found in the data, where Sprenger et al. (2014b) showed that positive events sentiment is more pronounced while X. Zhang et al. (2011) wrote that negative sentiment combination is more significant compared to positive. There are many other findings in the studies worth mentioning. For example, one study found that Twitter volume is negatively correlated with returns and positively correlated with volatility (X. Zhang et al., 2011), while sentiment is more correlated with stock trading volume than returns (Ruiz et al., 2012). Sentiment disagreement is a common measure in the studies as well, which also showed significant predictive power of trading volumes (Sprenger et al., 2014a).

2.4 Our research questions

The relationship between Twitter and the stock market is still in an early stage but rapidly growing. The aim of this study is to contribute to this research field by testing found relationships in the literature with new data and methodologies as well as by adding distinct perspectives. Bellow we describe and discuss the three tested hypotheses in this paper.

H1: *Twitter sentiment has predictive power of S&P 500 index returns.*
This is naturally a common hypothesis in literature. The aim of this hypothesis is therefore to simply replicate some of the main results found in previous research with new data and methodologies, as well as to extend the research in several ways. In addition to the traditional search queries (i.e. S&P 500 and Bullishness), our data includes market terms (i.e. search terms as 'equities' and more) and company based tweets. Thus, with higher variation in search queries as well as with a larger extent of data, we want to verify and extend current research.

Additionally, within this research area, the construction of the sentiment datasets requires an extensive number of steps and the possible options are vast. Therefore, as Twitter is still a rather new medium, we can see a lot of experimentation in the methodology part. Consequently, we also want to add to the field by comparing and extending some methodologies within sentiment analysis and sentiment indices construction. The extensions are discussed throughout the section 4. However, the main methodology contribution in this study is a comparison of a traditional sentiment analysis tool (Financial Loughran and McDonald Lexicon - LM) and a newer Vader NLTK method. To our knowledge there are no well-cited paper that used Vader NLTK method for the tweets' sentiment analysis in the financial literature.

In relation to this hypothesis, we also construct a new intuitive sentiment index by taking advantage of our larger dataset. We construct this new sentiment index by aggregating the sentiment of the main individual companies in the same way as the S&P 500 composite is aggregated. Additionally, it could be naturally assumed that not only the public opinion

about the stock but also the public opinion about the company's products and services should have an effect on the returns (Pagolu, Reddy, Panda, & Majhi, 2016). Therefore, we construct a unique company-based index by focusing not only on financial tweets. To our knowledge, non-financial Twitter companies sentiment (without 'cashtag') has never been analyzed in the financial literature.

H2: *Returns of some industries are more sensitive to Twitter sentiment.*
Extending the research related to company based sentiment, we investigate whether some industries are more sensitive to changes in the public mood compared to others. We believe that some industries might be more prone to sentiment. The reasoning behind this hypothesis stems from the previous discussion about representation bias. It could be argued that investors have stronger prior irrational opinions about certain industries compared to others, and therefore in those situations they might trade more on sentiment rather than fundamental information. While on the contrary, industries with low prior investors opinion should be less affected by sentiment. To our knowledge, there is no studies in literature that checked the industry based sentiment effect on returns using Twitter data.

H3: *Twitter sentiment has stronger predictive power of returns for companies with a higher concentration of retail investors.*
It is conventionally believed that retail investors are more sensitive to sentiment compared to institutional investors. With this hypothesis we want to verify or deny this belief. Tetlock (2007) found that small stocks are more affected by sentiment and he argued that those results were driven by the fact that the individual investors are more sensitive to publics' mood swings. However, he did not directly test whether the retail (individual) investors share in the company is a significant factor in explaining the relationship. Therefore, with this hypothesis we want to substantiate that argument by directly testing it.

To summarize, we add to the literature by 1) trying to replicate the main previous findings with a newer and broader dataset; 2) testing the traditional financial LM sentiment analysis method against the Vader NLTK sentiment analysis method, 3) examining whether certain industries are more predictable than others, and 4) exploring whether the share of retail investors in a firm has an effect on its return predictability using sentiment.

3 Data

3.1 Twitter data

We have collected data of original public tweets from Twitter for the period 2017-01-01 - 2018-01-01. For each record we have a tweet identifier, user identifier, tweet text, date and time of submission, number of retweets and number of likes. The data was collected using search queries of the targeted companies, S&P 500 and the market terms (the list of used queries can be found in Table 7). The company related tweets were collected for the 102 biggest U.S. companies of S&P 500 composite. We collected company related data partly for testing the relationship between company-based sentiment and S&P 500 returns and we expect that the 102 biggest companies are sufficient to test this relationship, since S&P 100 and S&P 500 indices' returns have daily correlation of 98.1% in the tested sample. In addition, we decided to use the search queries of the entire name, rather than a 'cashtag' (common financial way in Twitter to reflect the company with a dollar sign ($) in front of the ticker (Daniel, Neves, & Horta, 2017; Y. Mao, Wei, Wang, & Liu, 2012; Smailović et al., 2013)), because it was a common recommendation for future studies in the literature (Daniel et al., 2017; Y. Mao et al., 2012). In addition, not only the public opinion about the stock but also the public opinion about the company and its products should have an impact on the returns (Pagolu et al., 2016). Furthermore, we have divided market terms in the words with implied bullishness (bearish / bullish), S&P 500 search query and general market words, such as 'equities', 'stock market' and similar. The discussion for chosen search queries and grouping follows in section 4.3.1, while the used data can be found in Table 7.

During and after the process of collecting tweets, we filtered the data (see Figure 1). Initially, we manually filtered the companies data and excluded the names that might refer to other general words (i.e. Apple, Oracle). We also eliminated companies that have undergone the merger during the year. Moreover, we filtered the data to contain only the English language texts, because the further sentiment analysis tools are mainly trained for English lexicons (used sentiment analysis tools are discussed in section 4.1). As discussed before, Twitter is one of the biggest social media outlets with ∼500 million newly generated tweets a day (Aslam, 2018). In order to eliminate the noise in the data and optimize the data accumulation process, we collected only the 'top tweets'. 'Top tweets' reflect the proprietary Twitter algorithm-based assessment of the most impactful tweets (McGee, 2010). After filtering the data, our database was composed of ∼95 million tweets. The companies are divided in 7 groups of industries, based on Global Industry Classification Standard (GICS): Healthcare, Financial, Industrial, Information Technology, Consumer Discretionary, Consumer Staple and Other industries. Other industry is composed of Energy, Telecommunication Services, Real Estate, Utilities and Materials (each companies classification can be found in Table 7). This industry based classification will be used in hypothesis 2 testing.

3.2 Financial data

We collected the daily closing prices of S&P 500 from the Yahoo!Finance (2018) database for the period 2017-01-01 - 2018-01-01. For the same period, we collected the prices of

each company in the sample from the Compustat database (2018). The daily prices were converted to log returns in line with previews research (H. Mao et al., 2011):

$$R_{t,i} = log\Big(\frac{S_{t,i}}{S_{t-1,i}}\Big) \tag{1}$$

where R_t is return at time t and S_t is the closing price of a financial asset i. The descriptive statistics of used variables' returns can be found in Table 3.

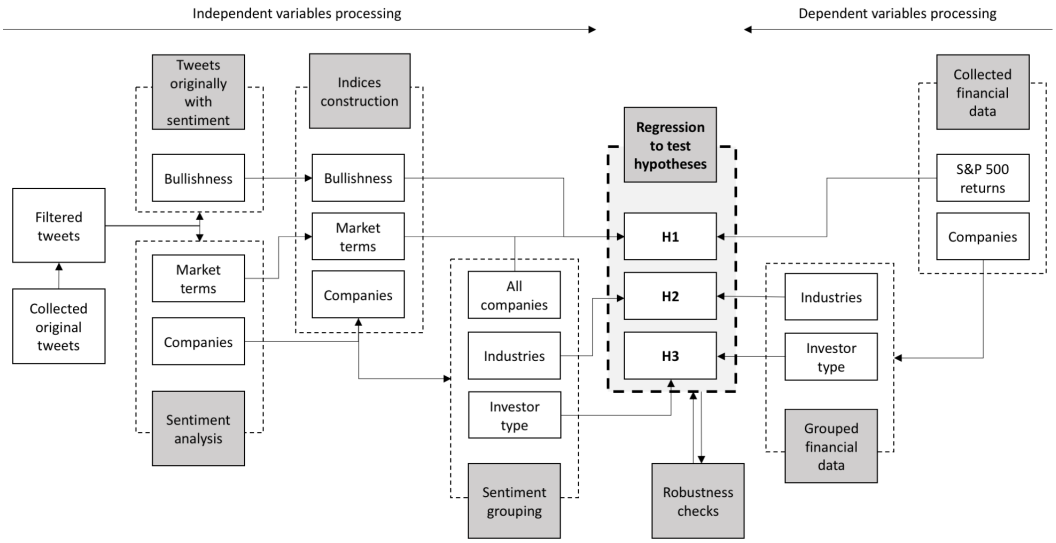
In addition, we collected the market capitalization data for each sample company from the Compustat database (2018), which is used later for data grouping (market-weighting). Trading volumes for S&P 500 as well as for each company were also collected from the Yahoo!Finance database (2018). The share of institutional investors and insiders were collected from Thompson Reuters EIKON (2018) for the day 2018-03-16. We calculated retail investor share using the following formula: $RetailHoldings_i = 1 - InstitutionalHoldings_i - InsiderHoldings_i$, where i is the company in consideration. We expect that the investor structure in the company does not change significantly over a one year period, therefore we did not include the time dimension. The companies with the highest and lowest retail investor share are marked in Table 7.

4 Methodology

The research methodology includes further processing of data, sentiment analysis, indices construction, grouping of dependent and independent variables and empirical relationship modeling. The steps of the process are depicted in the Figure 1

Figure 1: Steps of data collection and processing

The figure depicts the steps of Twitter and financial data processing done before the regressions. In the previous part we discussed the outer layers of the figure, including Collected original tweets, Filtered tweets and Collected financial data. The next step is Sentiment analysis. The methods used in Sentiment analysis are discussed in section 4.1. Bullishness has already implied sentiment, therefore the text analysis is not necessary. After the sentiment score is given to every tweet, the Indices construction is done. The methods used in this part are discussed in section 4.2. After the indices, we do Sentiment as well as financial data grouping (Grouped financial data). This part is also discussed in section 4.2. Then we move to construction of empirical models to test the hypotheses (Denoted in Figure as 'Regression to test hypotheses'), which is described in section 4.3. Finally each regression results are checked for robustness and necessary parts are transformed. The preformed robustness checks are discussed in part 4.4. Bullishness refers to tweets with search queries, such as bullish and bearish, market terms refers to both S&P 500 related search queries as well as general terms, such as 'equities' and 'stocks'. Companies refers to companies related tweets as well as financial information.



4.1 Sentiment analysis on Twitter data

Sentiment analysis is the method of processing natural language, using textual analysis and computational linguistics, that aims at identifying and extracting the opinion, subjectivity and emotion from the source text (Pang & Lee, 2008; Liu, 2012; Wilson, Wiebe, & Hoffmann, 2005). There are two most common areas in the literature for the textual sentiment classification: dictionary-based analysis/lexical and machine learning (Kearney & Liu, 2014). The dictionary-based technique maps the text at hand with a provided list of words. This approach depends highly on the quality of the dictionary and the weighting algorithm (Kearney & Liu, 2014). The machine learning technique classifies the text sentiment based on learned dynamics from the training data. We decided not to use the machine learning technique because 1) the tweet data we use does not have predefined sentiment for training, 2) the extensive level of complexity and 3) because machine learning algorithm does not have significant advantages over dictionary-based techniques for, specifically, social media data classification (Gilbert & Hutto, 2014). We decided to use two sentiment analysis tools

for analyzing texts of market and company related tweets: Vader NLTK (NLTK) as well as Financial Loughran and McDonald Lexicon (LM). The tweets with the predefined sentiment words, such as bullish and bearish, were not translated into sentiment, because they have already implied sentiment.

Among dictionary based techniques, there are no consensus on which one performs the best (Gilbert & Hutto, 2014; Ribeiro, Araújo, Gonçalves, Gonçalves, & Benevenuto, 2016; Araujo, Reis, Pereira, & Benevenuto, 2016; Lin et al., 2018). The results highly depend on the way the dictionary was created: which industry and which media outlet the algorithm was trained on and what training data size was used (Araujo et al., 2016; Lin et al., 2018). Vader NLTK[2] is a new and simple technique and was created for micro-blogging and social media sentiment identification, especially Twitter (Gilbert & Hutto, 2014). Many dictionary based techniques focus on identifying the number of positive and negative words in the text (Kearney & Liu, 2014). In addition to that, Vader NLTK also includes the strength of emotion (intensity) and sentence characteristics processing (Gilbert & Hutto, 2014; Ribeiro et al., 2016). To be more precise, Vader includes the predefined treatment of negation, punctuation, capitalization, constructive conjunctions (i.e. but) and strengthening adjectives (i.e. extremely good) (Ribeiro et al., 2016). Furthermore, Vader NLTK outperformed many well known dictionary-based techniques as well as machine learning models, especially in English tweets classification (Gilbert & Hutto, 2014; Ribeiro et al., 2016; Araujo et al., 2016). We know that some of the financial papers used several features of Natural Language Toolkit, however, to our knowledge there are no well-cited paper that used Vader NLTK for the tweets' sentiment analysis in financial modeling.

We used Vader NLTK compound score, which includes normalization and the intensity of the sentiment and ranges from -1 (extreme negative) to 1 (extreme positive), while values around 0 represent neutral sentiment. Some examples of tweets and their sentiment scores can be seen in Table 2. Some literature suggests to use this convention: score ≥ 0.5: positive; $-0.5 >$ score < 0.5: neutral; score ≤ -0.5: negative (Lin et al., 2018; Gilbert & Hutto, 2014), however, we did not find added value of this convention. The reason might be the limitation of the data: by using conservative thresholds, we decrease the amount of data points and fail to explain smaller variations in returns. The NLTK method has high positivity in our data sample: 70.24% of tweets with identified sentiment during the trading day were classified as positive (see Table 7).

Due to our finance focus, we also used the Loughran and McDonald Lexicon (LM list) for the sentiment analysis. LM list expanded GI/Harvard negative words list with finance-specific words: the LM lexicon was created based on large sample of 10-Ks fillings with a clear direction of analyzing financial texts (Loughran & McDonald, 2016; Kearney & Liu, 2014). Traditional dictionaries misclassify words as negative 73.8% of times while they are not negative in finance context(Loughran & McDonald, 2011), therefore a finance-based dictionary should increase the accuracy in analyzing tweets in relation to companies and financial markets. The LM lexicon outperformed several known methods in classifying data

[2]Vader sentiment analysis was recently added to Natural Language Toolkit, also known as NLTK. However, in this paper we use NLTK only in reference to Vader itself.

on company, industry and index level(Li, Xie, Chen, Wang, & Deng, 2014). In addition, LM is manually constructed, which is more accurate compared to automatic or semi-automatic lexicons (Li et al., 2014). The LM lexicon is also one of the most common method used in financial literature (Cortis et al., 2017; H. Mao et al., 2011) and it's popularity is only growing (Kearney & Liu, 2014). The disadvantage of LM is the fact that it was trained on long texts rather than micro-blogs and thus it might not be optimal in analyzing tweet texts. We used the common way of calculating sentiment polarity (Twedt & Rees, 2012; Kearney & Liu, 2014):

$$LM_i = \frac{N_{i,positive} - N_{i,negative}}{N_{i,positive} + N_{i,negative}} \tag{2}$$

where LM_i stands for LM sentiment polarity score of one tweet (i), $N_{i,positive}$ and $N_{i,negative}$ refer to a number of positive and negative words in a tweet text respectively. Similarly to NLTK, the score ranges from -1 to 1, while 0 being neutral. See the examples in Table 2. We used the updated 2016 master LM list for the sentiment analysis. The positivity was lower compared to NLTK method: 46.88% of classified tweets during trading time were positive (see Table 7). In addition, the NLTK method classified more tweets than the LM method during trading time as having a non-neutral sentiment. The conservative measure of LM might be a limitation in analyzing the relationship between the classified sentiment and returns.

Table 2: Examples of tweets and their sentiment scores

The table shows the randomly selected tweet texts and computed Vader NLTK compound scores (NLTK) as well as the polarity scores using Loughran and McDonald Lexicon analysis (LM). Both sentiment scores range from -1 and 1, where -1 represents negative sentiment, 1 - positive, and scores around 0 represents neutral opinion.

		Sentiment score	
#	Tweet text	NLTK	LM
1	"Trade liquid stocks from S&P 500 long and short positions up to 40% on profits"	0.440	0.333
2	"I didn't know that S&P 500 dropped 3% in the last three Januaries"	0.000	-1.000
3	"While the S&P500 closed at an all-time high last week volume has not been especially convincing ..."	-0.356	0.000
4	"I just reviewed the SP500 index and all of my investment charts. Feel - it is time to put out a warning of an SP500 and NY A/D"	-0.340	-1.000
5	"Learn strategies proven to beat the S&P 500"	0.000	0.000

4.2 Sentiment conversion to indices

After we have converted every individual tweet to a sentiment score (s_t)[3], the scores were aggregated to form sentiment indices (I_t). To do so, one needs to define the time thresholds (T to T-1) as well as the aggregation formula.

[3]Actually, every tweet has two sentiment scores as the NLTK and LM method were used, except for the Bearish and Bullish keywords, thus we have done the remaining steps with both scores separately in order to be able to compare the two methods, but due to notational convenience we do not denote this with any subscripts going forward.

Regarding the agregation formula, Antweiler and Frank (2004) conducted a deeper discussion in this area and proposed three different formulas. We decided to use two out of the ones suggested that are presented as equation 3 and 4 because we found them more appropriate for our research. The first equation presented in this section is neutral to the number of tweets (count-neutral), while the second one is count-dependent that leads to the sentiment index being amplified with the number of tweets that day. Antweiler and Frank (2004) found similar results between the methods but equation 4 was the slightly better performer. This result is also consistent with previous research which found that number of messages had an effect on the results (H. Mao et al., 2011)[4]. Equation 4 is commonly used in the literature (H. Mao et al., 2015; Sprenger et al., 2014a).

Daily count neutral sentiment index:

$$I_{t,i} \equiv \frac{M_{t,i}^{BUY} - M_{t,i}^{SELL}}{M_{t,i}^{BUY} + M_{t,i}^{SELL}} \tag{3}$$

Daily count dependent sentiment index:

$$I_{t,i}^{*} \equiv \ln\left(\frac{1 + M_{t,i}^{BUY}}{1 + M_{t,i}^{SELL}}\right) \tag{4}$$

$$M_{t,i}^{c} \equiv \sum w_i \dot{|} s_{t,i}^{c} | \tag{5}$$

for the two methods, M^c is defined as in equation 5 where $c \in \{BUY, HOLD, SELL\}$ indicates the sentiment category of a tweet that was classified by the sign of the sentiment score, i.e. a positive, neutral or negative sentiment score represents a BUY, $HOLD$ and $SELL$, respectively. This is a very commonly used classification method in the literature. Furthermore, the weight w_i is defined as normalizer, calculated as $1/|s_t|$ that is in line with most literature. In this case for example $M_{i,t}^{BUY}$ would simply be the number of tweets with a positive sentiment score during time t with the keyword or keywords i that we want to build an index I for. [5] Lastly, one could think that weighting the sentiment scores by the likes and retweets from the corresponding tweet might provide additional information. However, a study that analyzed individual tweet messages found that tweets with higher quality information was not retweeted more often (Sprenger et al., 2014a). In addition, the timing of likes and retweets are lagging the tweets themselves. Thus, as we want our results to be interesting for real time usage, we do not use such weighting system in line

[4] We investigated this equation further, and we could see that this formula was not so 'count dependent' for search queries with more tweets. As many of our queries had a substantial number of tweets, equation 4 may be closer to the Neutral measure than anticipated.

[5] We also investigated additional weighting methods: threshold-based classification and equally weighted classification. As discussed before, some studies suggest to use thresholds for assigning the sentiment to positive and negative texts: we used polarity of $+/-0.2$ as well as more conservative $+/-0.5$ as thresholds for positive and negative text classification, respectively. None of these two methods added any value to our results. We also investigated another weighting method i.e. an equally weighted score, since we wanted to capture the intensity of the sentiment. We expected, especially, that Vader NLTK method would show some different results, due to its structure. However, we did not see any significant differences. We concluded to use the normalized weighting system due to its popularity among researchers. In this paper, we only report the results for normalized method but upon request other results can be presented as well.

with literature (Antweiler & Frank, 2004).

Regarding the definition of T and T-1, we simply set all tweets posted during trading time as T and all tweets posted during previous trading time as T-1. Trading time is defined as 9:30 AM - 4:00 PM (GMT-04:00) during weekdays and excluding the weekends and holidays. The literature is not consistent in this distinction. The key reason for our choice is to use the most relevant tweet data and control for irrelevant information outside the trading hours. The time outside trading hours represent around 80% of total hours a year. We expect that tweets during those hours have a smaller impact on the trading behavior and including a proportionally large amount of them might just add noise to the dataset.

As a note, research (Tetlock, 2007) has found that negative sentiment has more effect than positive one [6]. The indices we use neutralize the distinction between the two and only focus on the change in the positivity. We can see the risk of using these indices since they do not take into account the relative importance of negative versus positive tweets. However, in oder to be able to compare the results with other studies we still use the previously mentioned index building formulas.

Sentiment indexes are built for every search query, for example *S&P 500*, *bullishness*, *market terms* as well as for each individual company as in equation 3 and 4. However, for companies there was an additional step where they were aggregated into one sentiment index using an equation 6. Thus, we have four different indexes for count-dependent and four different indexes for count-neutral measures at this stage for testing our first hypothesis[7]. The indices are graphically represented in Figure 3.

Daily aggregated company sentiment index combining all the 102 companies:

$$I_t^{Aggregated} \equiv \frac{1}{\sum_{i=1}^{102} w_{i,t}} \sum_{i=1}^{102} w_{i,t} \cdot I_{i,t} \tag{6}$$

where $I_{i,t}$ is the sentiment index value and $w_{i,t}$ is the market capitalization for company i at time t. The companies were aggregated in this manner (market capitalization-weighted) in order to replicate the way market returns for S&P 500 are calculated.

To test the hypothesis concerning industries, equation 6 was used again but only aggregating the companies belonging to a specific industry to create their corresponding sentiment index. The dependent variable in this case, which is used in regressions presented later, is market weighted returns comprising of the companies belonging to a specific industry, i.e. same companies are used in the industry sentiment index as well as for industry market weighted returns. In addition, we also used equally weighting for industry grouping to avoid the dependency of the company size. To the similar convention, we constructed the indices for Investor Type (H3). We used equation 6 to group the sentiment of companies with high

[6] These results were not confirmed using Twitter medium.

[7] Actually, we have twice as much indices, since we construct them both for LM as well as NLTK sentiment scores.

retail investors share and low retail investors share. The dependent variable used the same market weighting principle. The number of companies grouped by industry can be checked in Table 3 and the names of the companies can be found in Table 7 for the industries as well as for the investor types. For H2, the firms were grouped in accordance with the broad Global Industry Classification Standard (GICS). A more detailed sub-grouping was not done in order to avoid small sample problems. A small sample might lead to non representative results for the industry at hand and include high dependency on one big company.

4.3 Empirical methods of analysis

After we have created the indices, it is finally time to use them for testing their relationship with the stock market returns. Kearney and Liu (2014) discuss the different methods that are used to model the relationship between the sentiment and asset returns. The most common method is a linear regression (autoregressive distributed lag model - ADL) with addition of the most common control variables. Some studies also use VAR models as well as the less common panel regressions. Chen and Lazer (2013) discuss some reasons why the regression model is used and it comes down to the speed of it when applying trading strategies in real time specifically when using large amount of data as twitter data. They also argue that it provides benefits over commonly used classifier, since it give the indication of level not just a direction. The VAR model is usually applied to see the interrelationship in both (all) directions between the two (all) variables of choice. However, as Brown and Cliff (2014) found evidence of both way directional relationships as well as other studies found predictive power of the stock market using twitter data (Bollen et al., 2011; H. Mao et al., 2015), we will focus directly on the directional relationship concerning our hypotheses testing, i.e. we want to see the predictability of returns with twitter sentiment as a main explanatory variable. For this reason we use the most common method at this stage of the methodology: autoregressive distributed lag model with control variables. Its general form is presented in equation 7.

$$R_t = \alpha + \sum_{j=1}^{n} \beta_j^R R_{t-j} + \sum_{j=0}^{n} \beta_j^S S_{t-j} + \sum_{c=1}^{C} \sum_{j=0}^{n} \beta_j^c X_{t-j,c} \tag{7}$$

where R_t, S_t, and $X_{t,c}$ are the return, sentiment index value, and vector of control variables c on day t, respectively. Some studies do not use the contemporaneous variable, however, research found significant intraday predictability (Antweiler & Frank, 2004), thus we do not want to omit this variable even though its coefficient may reflect a two sided relationship.

Concerning control variables, apart from the lagged returns that are commonly included to control for momentum or autocorrelation, there are no consistent variables in the literature. This is expected due to the fact that the daily returns predictability has high complexity, and the studies that find some explanatory significant variables become the basis for trading strategies that may 'correct' asset prices and the significance of the variables could thus fade away. Some studies do not use any additional control variables after controlling for

autocorrelation in returns. That being said, we follow studies that control for liquidity effects as well (H. Mao et al., 2015; Tetlock, 2007). For the industry and investor type based regressions we use the market return as an additional control variable. It is also in line with other studies (Antweiler & Frank, 2004).

4.3.1 Rationalization for the chosen main explanatory variables in H1

In the first regression we use a sentiment index built only with tweets containing the keyword with the same name as the financial index we are predicting returns for, namely the keyword S&P 500. This is the most common method in the literature not only for predicting S&P 500 returns but also for predicting other indexes as well as individual stocks (Si et al., 2013; Sprenger et al., 2014a, 2014b; Ruiz et al., 2012; Smailović et al., 2013). Consequently, we start with this regression to simply see if we can find similar results as previous research with new data and methodology.

Secondly, we use keywords bearish, bear market, bullish, and bull market to form another independent variable, due to multiple reasons. It does not rely on any natural language processing technique as the words already have a predefined sentiment within them. Thus, this tests our results for robustness in a small way, however, the sentiment signal is different in its fundamental characteristics as well. For example, we expect there to be higher degree of autocorrelation with this signal as it is more forward looking feeling of the markets rather than instant reactions to variations in key factors. We expect the bullishness view to shift slower than the S&P 500 sentiment index. Lastly, we also see it as some kind of "improved" version of the very well known study (Bollen et al., 2011), where they only used tweets that contained, the words 'I feel', 'I have a feeling' and similar. The reason we see it as a similar measure in the first place comes down to that all these keywords do a good job of maximizing tweets that contain actual sentiment or feelings, and thus also filtering out neutral tweets and plain facts or similar potential noise. Furthermore, it is an improvement due to the closer connection to the financial markets. Thus, we see this being a more clean sentiment signal with less noise for our usage, with the forward looking distinction. This variable has also been used in literature (H. Mao et al., 2015).

As the third main independent variable, we use market terms related to the general stock markets. Studies show that combing keywords can improve prediction accuracy (X. Zhang et al., 2011). Thus, this variable could improve prediction accuracy. In comparison to the S&P 500 signal discussed before, we find this variable to be broader. However, it is still related as the S&P 500 is commonly referred to as the main stock market performance indicator. Regarding autocorrelation, we expect it to be in between the previous mentioned sentiment signals as its broader than the first (S&P 500), but not as forward looking as the second (bullishness).

The forth and last sentiment signal, related to hypothesis one, is constructed in a completely different but intuitive way by aggregating the sentiment signals from the main individual companies belonging to the S&P 500. Beside it to be different in its construction, it may also capture sentiment from a completely different audience. For the first three sentiment

indices we expect the tweets to be mainly financially related. For the companies, we can expect that the tweets are more related to the companies' products, services or latest news, since we did not use the financial tweets convention of using a 'cashtag'. We also expect the highest autocorrelation with the reasoning that companies cannot change their products or services that often nor does the public opinion about such offers change fast and frequently.

4.4 Robustness checks

To test our results for robustness, we mainly make sure that the models are used in a correct way by checking if the needed assumptions are satisfied. Before running the regressions we have tested all the variables for stationarity and multicollinearity. After the regressions we also check the residuals for heteroscedasticity and autocorrelation.

The information about not met assumptions we state in the description part of the regression results. But in summary, we found that some independent variables are trend-stationary which we detrended. Regarding multicollinearity, we see no significant issues in any of regressions, since we are not close to 80% correlation (used rule of thumb threshold). Further, we control the residuals for autocorrelation using the Durbin-Whatson test and we did not receive any red flags from this test (All presented regression results are within range of the test statistic from 1.5 to 2.5). Lastly, we found some heteroskedasticity in our residuals based on at least one of the used tests (we used White Test and Breusch–Pagan test). This implies that the coefficients are unbiased but that the variance of those could either be under or overstated, affecting the p-values. Due to this, we recalculated the p-values using MacKinnon and White's (1985) alternative heteroskedasticity robust standard errors. Specifically, we used the HC3 method as proposed by research when dealing with a sample size of 250 or less observations (Long & Ervin, 2000).

5 Empirical results and discussion

In this section, we present the results of our research. We start with analyzing descriptive statistics of used variables and then we move to discussing the results of the regressions. We discuss both the results found of different methodologies as well as the results of hypotheses. Finally, we end this section with the discussion of our models' limitations. For the reader's convenience we only mention the failed robustness checks and done transformations in the regression table descriptions. In addition, we limit the discussion of results in relation to data drawbacks. We acknowledge that this is highly possible explanation, however, discussing it in every part of the text might become cumbersome.

5.1 Descriptive data analysis

In Table 3, we show the descriptive statistics of each variable used in the upcoming regressions, for H1, H2 and H3. We have categorized the statistics in the order as it will be discussed: 1) pure tweets related statistics that only concerns the sentiment indices, 2) index composition statistics that are the same for the sentiment indices as for the return indices, and 3) general descriptive statistics that is unique for each regression variable.

1) Firstly, the number of tweets collected for the first four explanatory variables used in H1 (sentiment indices for S&P 500, Bullishness, MarketTerms, and Companies) are very different. This is naturally due to some sentiment indices having more search queries. For example, S&P 500 has around 28 thousand classified tweets during trading hours while the companies' sentiment index has around 15 million tweets as it contains data from many different companies. Surely, the amount of tweets depends not only on a number of search queries used, but also the popularity and awareness. In addition, more tweets should improve the true sentiment signal and thus making it less prone to noise (reduce the impact of outliers). However, in our sample, 27 thousand is the lowest number of tweets collected, which is large enough compared to previous studies. The number of tweets per company is more important than the total amount. For each group (both industry based as well as investor type based), the average number of tweets per company varies from 11.5 thousand tweets per company (Healthcare) to 450.5 thousand per company (Consumer discretionary), while 144.1 is an average among all companies. This could thus imply that the sentiment index from some industries (Healthcare) are more prone to noise than others. However, it is important to note that this may not be a full story. There are many possible explanations and interpretations. For example, one could assume, that the fewer tweets from the certain industries could be due to lack of general public interest in the company and thus more concentrated within industry experts, and therefore may have a larger influence on investment decisions. In this regard, the relevance of the sentiment may be larger. The noise created from not having the perfect sentiment method may still be larger for the industries with fewer tweets per company. This discussion is also relevant for the investor type variables used in H3. We can see that the Top15Retail variable was constructed using a significantly larger number of tweets compared to the Lowest15Retail variable. As the companies contained in the Top15Retail are more likely to be well-known, the firms with fewer tweets are more likely to be tweeted by industry experts. This conclusion also assumes that the share

of experts is higher for less known companies, which is reasonable. The main take away is that sentiment indices built with fewer tweets per company could be of more importance for predicting investor behavior.

Table 3: Descriptive statistics

The table shows all variables (independent, dependent and control) used in the following regressions (Reg_1-Reg_14). We divide the variables in Sentiment variables, Log returns variables and Trading Volumes. All presented sentiment variables (except cBullishness) are count-dependent sentiment indices made of NLTK scores. cBullishness is a count-dependent sentiment index made of 'bullish' and 'bearish' Twitter data (predefined sentiment Tweets - no sentiment conversion needed). cNLTK_S&P500 uses the tweets data that was collected using S&P 500-related search queries and cNLTK_MarketTerms variable refers to collected data of general financial market terms (the list of words can be found in Table 7). cNLTK_CompaniesIndex, each Ind_SentIndex and each InvType_SentIndex are the market capitalization-weighted sentiment indices comprised of all the individual company sentiment indices related to the relative group. The industries are: Health - healthcare, Fin - financial, Indust - industrials, IT - information technology, CD - consumer discretionary, CS - consumer staples and Other. Top15Retail represents a variable of the 15 companies in our sample with the biggest share of retail investors, while Lowest15Retail represents the 15 companies with the biggest share of institutional investors and insiders (lowest retail investors share). A Log return for an industry and investor type is the market cap weighted sum of log returns of the related individual companies. TradingVolume: S&P 500 is trading volume of S&P 500, while every other Trading Volume variable is a mcap-weighted sum of the trading volumes of the related companies. The related companies are either grouped by industry or by composition of investor type (Ind_TradingVolume and InvType_TradingVolume). All variables' statistics are calculated on the observations used in the regressions, except Trading Volumes are scaled up by multiplying the observations with 10^3 for visual purposes. The Reg column shows in which regressions the variables were used. Statistics are only presented for the variables used in the main regressions (not the ones in Appendix). tr_tweets refers to the total tweet number over trading time (excluding neutral tweets) used in constructing the indices. Pos shows the average of positivity among the sentiment measures. Detailed distribution of the positivity measure can be seen in Table 7. # represents the number of companies used to make the sentiment indices. Retail shows the average share of retail investors in the companies. Size refers to the average market cap level per company in the specific industry or group (expressed in USD bn). General descriptive statistics, or to be more precise: Mean, Standard Deviation (Std), Maximum value (Max) and Minimum value (Min), are calculated for each variable used.

		Tweets data		Index composition			General descriptive statistics			
Variable	Reg	tr_tweets	Pos	#	Retail	Size	Mean	Std	Max	Min
Sentiment variables:										
cNLTK_S&P500	1,5	27,745	63%				0.5419	0.3769	1.6809	-0.7340
cBullishness	2,5	231,046	74%				1.0315	0.2099	1.5618	0.2101
cNLTK_MarketTerms	3,5	616,646	62%				0.5378	0.2407	1.0662	-0.2066
cNLTK_CompaniesIndex	4	14,699,509	70%	102	21%	118	0.8400	0.0642	1.0294	0.6396
Ind_SentIndex: Health	6	207,514	67%	18	18%	115	0.7803	0.1850	1.2913	0.2838
Ind_SentIndex: Fin	7	438,096	72%	15	19%	114	0.8164	0.1446	1.1806	0.4278
Ind_SentIndex: Indust	8	579,250	74%	14	23%	82	1.0473	0.1516	1.4160	0.4600
Ind_SentIndex: IT	9	6,314,022	73%	18	18%	133	0.9683	0.1118	1.2190	0.6326
Ind_SentIndex: CD	10	5,406,480	69%	12	21%	134	0.9202	0.1370	1.2045	0.2603
Ind_SentIndex: CS	11	1,090,789	68%	12	25%	133	0.7446	0.1425	1.1159	0.3567
Ind_SentIndex: Other	12	663,358	70%	13	26%	117	0.6040	0.1486	0.9746	0.2600
InvType_SentIndex: Top15Retail	13	2,092,518	72%	15	38%	154	0.8190	0.1142	1.1793	0.4301
InvType_SentIndex: Lowest15Retail	14	269,420	76%	15	7%	76	0.9930	0.1867	1.6182	0.4014
Log returns variables:										
S&P 500 log returns	All						0.0007	0.0042	0.0136	-0.0183
Log returns: Health	6			18	18%	115	0.0005	0.0055	0.0169	-0.0136
Log returns: Fin	7			15	19%	114	0.0007	0.0095	0.0296	-0.0384
Log returns: Indust	8			14	23%	82	0.0005	0.0058	0.0148	-0.0207
Log returns: IT	9			18	18%	133	0.0012	0.0074	0.0335	-0.0288
Log returns: CD	10			12	21%	134	0.0006	0.0072	0.0360	-0.0733
Log returns: CS	11			12	25%	133	0.0004	0.0050	0.0225	-0.0148
Log returns: Other	12			13	26%	117	-0.0001	0.0054	0.0191	-0.0147
Log returns: Top15Retail	13			15	38%	154	-0.0002	0.0047	0.0172	-0.0146
Log returns: Lowest15Retail	14			15	7%	76	0.0011	0.0057	0.0174	-0.0232
Trading Volumes:*										
TradingVolume: S&P 500	1,2,3,4,5						3.4109	0.5174	5.7239	1.3498
Ind_TradingVolume: Health	6			18	18%	115	0.0066	0.0018	0.0143	0.0023
Ind_TradingVolume: Fin	7			15	19%	114	0.0206	0.0064	0.0559	0.0047
Ind_TradingVolume: Indust	8			14	23%	82	0.0107	0.0053	0.0465	0.0042
Ind_TradingVolume: IT	9			18	18%	133	0.0138	0.0038	0.0319	0.0053
Ind_TradingVolume: CD	10			12	21%	134	0.0077	0.0020	0.0162	0.0036
Ind_TradingVolume: CS	11			12	25%	133	0.0065	0.0018	0.0194	0.0025
Ind_TradingVolume: Other	12			13	26%	117	0.0104	0.0029	0.0208	0.0036
InvType_TradingVolume: Top15Retail	13			15	38%	154	0.0141	0.0040	0.0340	0.0050
InvType_TradingVolume: Lowest15Retail	14			15	7%	76	0.0027	0.0006	0.0052	0.0010

*The descriptive statistics of Trading Volumes were calculated using the data used in the Regressions * 10^3 for visual purposes

2) Secondly, we consider the index composition statistics. For the industry-based indices (both sentiment and return indices), a similar number of companies is in each category, ranging from 12 to 18. The similar distribution of companies should limit the problems arising from uneven grouping. Additionally, there is no big variation in the share of retail investors among the industries. Thus, the results of investor type based regressions should not be biased towards the industries. Naturally, the difference in average retail share in Top15Retail and Lowest15Retail is high. Further, we can see that the average market capitalization per company is rather similar among the industries, except for the Industrial industry that is notably lower. However, the variation among companies within the industries are big, for example, within the IT industry the smallest company had an average (over sample year) market capitalization of USD 42.4 bn, while the biggest - USD 555.5 bn. Therefore, market cap weighted variables such as Log returns, Trading Volumes and some Sentiment Indices, will be more driven by the largest players. Lastly, Top15Retail companies are on average twice as large compared to the Lowest15Retail companies. Since all companies in our sample are big, this might be explained by the fact that the bigger the company, the more well-known it might be to the general public and thus more popular for retail investors. If our company sample would have included smaller firms, this explanation might not hold, since some institutional investors might have restrictions to invest in small companies and thus the retail share would be higher in those.

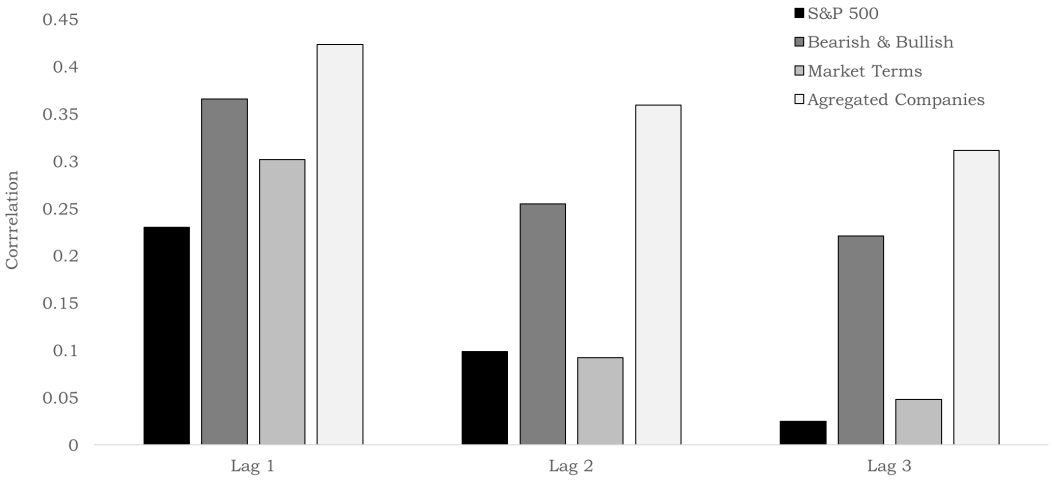
3) For the general descriptive statistics by specifically looking at the first four sentiment index variables, we see a lot of interesting findings. In the methodology we discussed why these signals were chosen and why they may be different in their characteristics. The variable statistics can be found in Table 3, but due to the importance of these variables and to give the reader and intuitive understanding, they are also graphically presented in Figure 3 (Appendix) together with their autocorrelation properties in Figure 2.

We see that the S&P 500 sentiment index[8] is the most volatile, closest to negative territory as well as it is the index with the least autocorrelation i.e. its values are less correlated with its past values. This is in line with our expectations and thus may support our discussion that this sentiment is mainly short term driven as it is the case for the day to day returns of the S&P 500. Further, we also expected Bullishness to be the most forward-looking sentiment and thus least likely to change as quickly. As the autocorrelation for this index is among the highest and persistent it may again be an indication that our discussion is in the right direction. Regarding the third variable, the sentiment index of the Market Terms, we see that it is not as volatile as S&P 500. We also mentioned that we believed it to be a broader measure but not as forward looking as bullishness. This could again be supported by the displayed autocorrelation, volatility as well as the level being lower. Lastly, the company index seems to be the most persistent as confirmed by the low volatility and the strongest autocorrelation. This is in line with our discussion that this measure could be mostly targeted towards a company's products and services from a non-financial perspective. Because neither the products nor the opinions change that quickly it would be natural for the signal to be the most persistent as we see. However, we cannot conclude this with

[8]S&P 500 sentiment index has trend stationarity, and in regressions we use detrended values, however, here we show the results before detrending for comparison reasons.

any certainty as this result may be mostly driven by the construction of the index. Namely, the companies index is an aggregated signal, thus it will neutralize the individual company sentiment indices to a large extent as they are combined. This effect can be partly seen through the volatility of industry sentiment indices. We can see that they range from 0.11 to 0.19, while cNLTK_CompaniesIndex has volatility of only 0.06. By combining companies into sentiment index, we neutralize the sentiment volatility. Another observation is that there is significant variation among average sentiment of the industries.

Figure 2: Autocorrelation for the the main explanatory variables
Autocorrelation for the fours sentiment indices from 1 to 4 lags, S&P 500, Bullishness, Market Terms and aggregated companies.



If we focus on returns, the volatility of S&P 500 is the lowest among the presented groups in Table 3. It is a combination of the chosen sample year as well as it is composed of more firms than the other variables. The most profitable was the IT industry in our sample with average daily return of 0.12%. The returns and volatility are different among the industries and could refer to the industry specific risks. In addition, there is a big difference between the average returns of Lowest15Retail and Top15Retail. However the return volatility is not that different, while the sentiment volatility for Lowest15Retail is significantly higher.

From the Trading Volumes part in Table 3, we can also see the variation in liquidity among the companies. Naturally, S&P 500 trading volume is highest since it is not divided by the number of companies, thus not comparable to the other groups. We can see that the Financial industry is the most liquid group in our sample with the highest average trading volume, while the Consumer Staple companies' are the least liquid. Since we market-weighted the Trading Volumes, the presented numbers might also be leaning towards the liquidity of the biggest companies in the industry. We can also see that the Lowest15Retail has a substantially lower liquidity compared to Top15Retail. This is expected because the Top15Reatil also has larger market capitalization.

5.2 Relationship between S&P 500 returns and Twitter sentiment

In this part we discuss the results of Hypothesis 1. We structured this part in two sections. In the first one we discus the results of different methodologies. We conclude that NLTK is a better suited sentiment analysis tool in this case compared to LM and that count-dependent indices construction technique carries several benefits over count-neutral. In addition, we discuss the lag selection method and conclude that 2 lags are optimal going forward. Therefore, in the following discussed results we use count-dependent NLTK sentiment indices and 2 lags for every variable at hand. We discuss the implications of using these methods and possible reasons of results in more details in subsequent section. After that, in section 5.2.2, we start discussing the results of regressions that model the relationship between S&P 500 returns and Twitter sentiment.

5.2.1 Methodologies related results

In the methodology section we discussed several ways of conducting the sentiment analysis and constructing the sentiment indices. Before analyzing the results of the specific regressions, we discuss which method of sentiment analysis works better comparing the performance of LM based sentiment and NLTK based sentiment as well as comparing count-neutral indices performance versus count-dependent indices performance (see sections 4.1 and 4.2 for clarifications). After that we discuss selected lag specifications for the following regressions.

Tables 4, 8, 9 and 10 report the results of the regressions between S&P 500 returns and sentiment. From the initial view, we can see that the regressions using NLTK based sentiment find more significant relationship than using LM method: the adjusted R-squared measures are higher and coefficients of independent variables are more significant (see Table 4 vs. 9 and 8 vs. 10) . This might be an indication that either NLTK is a better sentiment analysis method for the sample data or, if LM method is actually correct, that the relationship is just weaker. However, Reg_2 in each table (4, 8, 9 and 10) indicates a strong relationship between the S&P 500 returns and the independent variable, while being not subject to the sentiment analysis tools (bearish/bullish tweets). This lets us to believe that NLTK might be a better measure in analyzing tweets sentiment. There are several implications from the results that NLTK performs better than LM. First, it might imply that sentiment analysis method's performance depends more on what type of data it was trained on rather than the type of lexicon. In addition, it shows that the predefined treatment of certain sentence characteristics play an important role for micro-blogging texts. Finally, conservative sentiment measures (low 'coverage') might provide less benefits than expected, since it might fail to identify some important trends especially for the search queries that had smaller amount of tweets (i.e. S&P 500, some companies). However, it is important to note that both measures lead to the same coefficient signs in almost every regression (see Table 4 vs. 9 and 8 vs. 10) and thus could be concluded that they both model similar relationships. Despite the popularity of LM in the financial literature, we believe using NLTK might provide more accurate results and thus we use it for all reported regressions moving forward. For future research, a potential improvement for the sentiment conversion could be to combine the two methods: use the Vader NLTK specifications with a financial lexicon.

In a similar way we compare the performance of the count-neutral and count-dependent aggregation formulas for forming indices (see equations 3 and 4). Comparing Tables 4 vs 8, we can see that the performance of both sentiment indices is very similar: coefficients' signs are the same for independent as well as control variables. In addition, parameters significance and adjusted R-squared measures are very similar, too. This is in line with research, which finds that both indices perform similarly (Antweiler & Frank, 2004). At the same time, it shows that the inclusion of the number of messages (count-dependent) does not have a significant impact on the results. There could be opposing reasons for the found results. It is possible that the volume of messages is less important compared to the sentiment polarity, however it does not go in line with literature, which found high significance of sentiment volume variable (H. Mao et al., 2015). It is also possible that the logarithmic normalization of the count eliminated the significance of it, especially for the variables with high number of tweets. Therefore, inclusion of the not normalized volume of messages might improve the regression results. However, this was not found by literature that compared the indices (Antweiler & Frank, 2004), therefore, we did not perform this analysis. Finally, the reason could be the limitations of our data. Due to the fact that we only collected 'top tweets', we are highly dependent on Twitter's algorithm. Thus, the volume of messages becomes sensitive to the design of Twitter's proprietary 'top tweets' algorithm as well as it becomes sensitive to the assumption that underlying rules of the algorithm did not change over time. Due to the fact that both sentiment indices performed similarly, we chose to use the more common method in the literature for the further analysis of our results - count-dependent.

Finally, there is a lack of agreement in the literature which lags are the best in predicting returns using sentiment. Due to this fact, we have chosen the number of lags by minimizing Akaike information criterion (AIC) up to five lags in Reg_1 from Table 4. We have found that two lags were optimal in the specification of the model, which are going to be used for all following regressions in this paper (for comparability reasons). We tried to have as small amount of lags as possible in order to have a parsimonious model. In addition, since we have only big companies in our sample, we expect that both, overreaction and reversal to fundamentals, should happen in a very short period, which goes in line with the literature (Hong & Stein, 1999). Therefore, we believe that 2 lags supposed to be optimal to model the relationship.

5.2.2 Results of S&P 500 regressions

To test the first hypothesis, whether S&P 500 returns are affected by Twitter expressed sentiment, we constructed several linear regressions as discussed in the methodology part. The summary results can be found in Table 4. In each regression we control for lagged effects of S&P 500 returns and Trading Volume[9]

[9]Trade Volume was scaled down by 10^{12} for easier reporting of results.

Table 4: H1: Regressions results with count-dependent NLTK

In this table we present the results of five different regressions that try to model the relationship between dependent variable (S&P 500) log returns and selected independent variables. We used distributed lag regression model on daily observations. Log returns are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of S&P 500. All regressions include constant, control variables of scaled daily S&P 500 Trading Volume and its lags as well as lagged dependent variable (S&P 500 log returns). Independent variables - cNLTK_S&P500 and cNLTK_MarketTerms - are count-dependent sentiment indices made of NLTK scores. cNLTK_S&P500 uses the tweets data that was collected using S&P 500-related search queries and cNLTK_MarketTerms variable refers to collected data of general financial market terms (the list of words can be found in Appendix). cBullishness is count-dependent index of 'bullish' and 'bearish' Twitter data (no additionally added sentiment). cNLTK_CompaniesIndex is the market capitalization-weighted index comprised of each of the 102 biggest S&P companies count-dependent sentiment indices of NLTK scores. Each independent variable has two lags. In addition, cNLTK_S&P500 was trend-stationary, which we corrected by detrending the time series with polynomial of order 1 (linear trend). We also corrected Reg_1 and Reg_3 for the heteroscedasticity in residuals with heteroscedasticity consistent standard errors (HC3), because at least one of the tests rejected the homoscedasticity hypothesis (White test and Breusch–Pagan test). The stars indicate the significance of coefficients.

		Regressions with dependent variable - S&P 500 log returns				
#	Independent variable	Reg_1	Reg_2	Reg_3	Reg_4	Reg_5
1	cNLTK_S&P500	0.0059 ***				0.0027 ***
	cNLTK_S&P500_Lag_1	-0.0004				0.0002
	cNLTK_S&P500_Lag_2	0.0004				0.0006
2	cBullishness		0.0110 ***			0.0064 ***
	cBullishness_Lag_1		-0.0037 ***			-0.0028 **
	cBullishness_Lag_2		-0.0010			0.0005
3	cNLTK_MarketTerms			0.0098 ***		0.0062 ***
	cNLTK_MarketTerms_Lag_1			-0.0025 *		-0.0010
	cNLTK_MarketTerms_Lag_2			-0.0011		-0.0006
4	cNLTK_CompaniesIndex				0.0087 **	
	cNLTK_CompaniesIndex_Lag_1				-0.0043	
	cNLTK_CompaniesIndex_Lag_2				0.0050	
5	TradingVolume	0.0901	-0.1169	0.0087	-0.0567	0.0382
	TradingVolume_Lag_1	1.2214 ***	1.1749 **	0.8258 **	1.0557 *	0.9847 **
	TradingVolume_Lag_2	-0.9209 *	-0.3950	-1.0463 **	-1.4666 ***	-0.5013
6	S&P500returns_Lag_1	-0.1983 ***	-0.1165 *	-0.1572 ***	-0.1283 **	-0.1870 ***
	S&P500returns_Lag_2	-0.1216	-0.0311	-0.0428	-0.0700	-0.1163 *
7	Constant	-0.0004	-0.0080 ***	-0.0019	-0.0055	-0.0076 ***
	# observations	250	250	250	250	250
	Adj. R-squared	0.2832	0.2871	0.3261	0.0493	0.4783

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

The first regression (Reg_1 in Table 4) follows the common practice in the literature, where we try to explain the market index returns with tweets sentiment data, which was collected by using the index name as a search query. Looking at the results, we can see that the S&P 500 sentiment (cNLTK_S&P500) has a significant contemporaneous coefficient at 1% level. The coefficient is positive, which indicates the expected relationship - positive tweets are correlated with positive returns, which might be an indication of overreaction. The coefficient value means that an increase in sentiment index by 1 unit raises the return over the same day by 0.59 percentage points (pp). From an econometric perspective, literature suggests several explanations for why not-lagged values of independent variables have significant effects: either that (a) the variable has a finer-grained effect than the data allows to investigate, or (b) the variables are misspecified and they are associated with distant lags not included in the model, or (c) the effect is legitimate, but to the extent where it is not possible to distinguish which variable causes the other from the regression used (Granger, 1969; Geweke, 1982). We find evidence in the literature that argument (a) is feasible in this situation, but in combination with (c). Studies have shown that sentiment has predictive

power within 15 minutes intervals (Antweiler & Frank, 2004). Despite the fact that this research was done using different medium than Twitter, the results could be expected to be similar. This supports argument (a). However, Brown and Cliff (2004) found that the relationship is two sided: both the sentiment has an effect on returns, as well as returns have an effect on sentiment, which is relating to the argument (c). At this point we also want to stress that, since studies have shown that sentiment has a predictive effect on returns, we do not investigate the relationship from both directions, but rather focus on the direction that is related to our interests and hypothesis: we focus on explaining S&P 500 returns only. As a result of the significance in discussed coefficient, we cannot reject the null hypothesis at this stage - that sentiment has no valuable information for the prediction of S&P 500 returns. The predictive power might be lying within shorter periods.

On the contrary to the contemporaneous sentiment variable, the coefficients of the lags are insignificant at 10% level. However, the first lag has a predicted coefficient sign - negative. This implies that the price reversal to the fundamental value might happen after one day. There are several possible explanations of this insignificance. First of all, it is possible that the reversal to fundamentals happen after longer period of time, which is not included in our regression. However, due to high popularity and liquidity of S&P 500 index, it is also possible that its prices became very efficient and reverse itself even faster. In relationship to this argument and taking into account the assumption (a) explained in the previous part, it is possible that part of the reversal already happened over the same day, which led to lags being insignificant. If this is true, then non-lagged cNLTK_S&P500 coefficient is also under-reported. Further research needs to be done using shorter intervals.

Adjusted R-square of the regression (Reg_1) indicates that the 28.32% of dependent variable's variation is explained by the specified variables. In addition, we see that all lags of Trading Volume are statistically significant at least at 10% level. The first lag has a positive sign, while the second - negative. The former might indicate that higher liquidity the previous day leads to higher demand for S&P 500 asset(s) and thus leads to higher prices. Only the first lag of dependent variable is statistically significant and has a negative sign.

As discussed in methodology, we try to explain the variations of S&P 500 returns using bullishness index in Reg_2. To summarize the motivation of this variable discussed in the methodology, one can say: it is most likely a less noisy sentiment signal due to no sentiment analysis needed and it is also more forward looking sentiment index compared to the one used in Reg_1. Looking at the results, we can see that both cBullishness and cBullishness_Lag_1 are significant at 1% level and have the expected coefficient sign. The results are somewhat consistent with literature (H. Mao et al., 2015), where the researchers found the reversal to the fundamentals in the upcoming days using the Bullishness as independent variable. It is also interesting to note that H. Mao et al. (2011) also found that bullishness lags have a significant effect on DJIA returns. They found that lags 1, 2, 5, and 6 are statistically significant. The first lag had a positive coefficient and the second lag had a negative, while we found negative sign of the first lag and no statistical significance on the second lag. If we assume that DJIA and S&P 500 indices move in a similar way and are effected by similar underlying factors, the differences between our found results and theirs

might indicate that market became more efficient and investors, to some degree, already trade on the sentiment based strategies. Therefore second lag reduced in the significance in our findings and the reversal to fundamentals already happens the next day (compared to two days lag before). It is also possible that these two indices have different underlying factors and that S&P 500 prices are more market efficient compared to DJIA. Therefore, the contemporaneous variable in Reg_2 can be interpreted in line with the discussion under Reg_1. At this point, we can reject our null hypothesis, which we could not completely do in the previous regression as discussed.

Adjusted R-squared is similar to the previously discussed regression. This indicates that the usage of simple data that does not require additional sentiment analysis might be at least as good at explaining the variations of S&P 500 returns. Also, control variables show similar relationship as discussed in Reg_1, however, the significance decreased.

The third regression focuses on finding a relationship between the general market terms and the S&P 500 returns. Since S&P 500 is commonly referred to be the market performance measure, the general market terms such as 'stock market', 'equities', 'stocks' and 'indexes' should carry a statistically significant sentiment. These market terms have a broader perspective than the S&P 500 sentiment but not as forward looking as Bullishness (the more detailed discussion can be found in methodology part). In the Reg_3 both the cNLTK_MarketTerms and cNLTK_MarketTerms_Lag_1 are significant at 1% and 10% level, respectively. Therefore, we cannot reject with confidence the null hypothesis at this stage. Despite this fact the parameters have the same signs as in Reg_1 and Reg_2, therefore the similar explanations and reasons can be used here, both for contemporaneous as well as lagged variable.

In addition, the R-squared is slightly higher for the Reg_3 compared to other regressions discussed previously. This might be an indication that to a small degree the general market sentiment trends are more predictive than the sentiment of the index itself (in Reg_1). Moreover, these results go in line with X. Zhang et al. (2011), who showed that it is beneficial to make combinations of several terms with sentiment. However, the higher R-squared could have happened also due to the differences in data collected: we have a larger amount of tweets data for the general market terms compared to S&P 500 search query. The bigger amount of tweets might have led to better calibration of the sentiment analysis, and thus we received better results. Looking at the control variables, we can see that coefficients have the same signs and similar significance in comparison to Reg_1. Only the first lag of of trading volume decreased in significance from 1% level to 5% level.

Forth, S&P 500 is still the index of its composites, therefore in the Reg_4 we tried to model the relationship between the sentiment of the top 102 biggest U.S. companies and the S&P 500 returns (Table 4). As discussed more extensively in the methodology, this is an aggregated sentiment with a couple of expected key differences in its nature compared to the variables used in previous regressions. Briefly put, it is a market weighted index with sentiment from tweets that most likely have the least 'financial' information and are more product focused. The coefficient of cNLTK_CompaniesIndex is significant at 5%

level, however the significance is lower than found in other regressions. Also the lags of the sentiment are not significant at 10% level. However the coefficient signs are the same as discussed in Reg_1. In addition, we can see that the significance of the regression fit decreases dramatically compared to the previously discussed regressions (Adj. R-square is 4.93%), while control variables have similar coefficient signs. The significance of first lag of TradingVolume decreases compared to previous regressions, while the significance of second lag increases. In addition, first lag of dependent variable is significant at 5% level.

There are many possible reasons why we find less significant results in this regression. First, it is possible that as discussed before, S&P 500 index is rather a reflection of the market, and thus market trend sentiments are more important in modeling the possible relationships. This would imply that the logic used in Reg_4 supposed to be reversed in nature: the sentiment of market trends should have led to the increase in demand of S&P 500, which in turn translated to the increase in demand for composite companies shares (since they are the components of the index). This relationship would lead to the situation where S&P 500 becomes the driver of companies' prices, rather than the result. Second, it is also possible that our company based data collection was limited. We have collected a lot of tweets for each company, but some companies had a larger amount of tweets compared to the others. The companies that had a smaller amount of tweets might have contributed to a poor sentiment accuracy. Another limitation of companies data is the possibility of irrelevant information. We have filtered the data and excluded the companies that have the names of general terms, however, some of the used companies might still have irrelevancies, i.e. Abbott term could reflect some people with that same last name, like Diane Abbott. Also, we only collected 102 companies of the S&P 500 and we had to exclude several big companies due to generic name, i.e. Apple. These big companies that we excluded might have had a significant part of explaining S&P 500 returns variations (since it is a market cap-weighted index). Third, it is possible that people overreact more to financial information, compared to product driven information. Since this index had possibly the lowest amount of financial information compared to other discussed sentiment indices, the investor behavior might not have been impacted as much. Also, publics' opinion to the products and services are less volatile than information about the financial markets. The less variation in this sentiment index (as discussed in the descriptive statistics) might have hindered the possibility to explain the variations in returns. Finally, it is possible that some type of companies within the sample are more sensitive to public sentiment than others. This goes in line with study conducted by Smailović et al. (2013), which found that certain stocks are easier to predict using sentiment compared to others. Therefore, when we combine them in one index, the significant sentiment variations might have been neutralized by the sentiment of hard to predict stocks. In order to test the possible differences of companies' sensitivity to sentiment, we group the companies in relation to industries and investor type. We expect that these groupings could provide clearer explanations and we test them in hypotheses 2 and 3.

Finally, in order to check general market trends' effect on S&P 500 returns, we combine the first three regressions' independent variables (see Reg_5 in Table 4). As discussed before, these variables have different characteristics, however they all have the broader focus than

an individual company-based sentiment. The reasoning for this regression goes in line with the literature, which reports the increase in significant results when using sentiment combinations (X. Zhang et al., 2011). We can see from results that each of the contemporaneous independent variable is significant at 1% level and can add additional information in explaining the returns. This is an additional indication that future literature should focus on shorter time intervals in explaining S&P 500 returns. It is important to note, that none of the independent variables have the correlation higher than 41%, thus, we assume that the model does not have multicollinearity problem[10]. The signs of contemporaneous variables' coefficients are positive and the explanation of them could be similar as used in discussing Reg_1. The only lag that stayed statistically significant in this regression is the Bullishness first lag (cBullishness_Lag_1). This goes in line with our discussion that Bullishness has more forward looking effect. The sign of this variable supports the assumption of reversal to asset's fundamental value. However, cNLTK_S&P500_Lag_1 coefficient sign unexpectedly changed to positive. The finer grained data is needed to explain this result.

This regression explains 47.83% of dependent variable daily variation. The signs of control variables stayed the same as in previous regressions. However, the significance changed. Only the first lag of TradingVolume is significant at 5% level, while both lags of dependent variable shows the significance of up to 10%.

5.3 Industry based returns and sentiment

In oder to test the hypothesis, whether some industries are more related to sentiment than others, we conducted seven separate regressions: one for each industry. The results can be seen in Table 5. The number of companies grouped by industry can be checked in Table 3 and the names of the companies can be found in Table 7.

From the regressions we can see that the contemporaneous S&P 500 return is a strongly significant variable in each regression. The coefficient of each regression indicates that there is a high positive correlation between this control variable and industry returns. The coefficient says that if the S&P stock index goes up with 1%, this would lead to a 0.47 and up to a 1.60 percentage-point increase in returns depending on the industry. In addition, compared to previous regressions in Table 4, the trading volume (control variable) decreases in predictive power and is no longer significant in any of the regressions.

[10] We use a common convention used in literature that identifies multicollinearity problem only if the correlation among independent variables is higher than 80%.

Table 5: H2: Market cap-weighted Industry Regressions results

In this table we present the results of seven regressions that try to model the relationship between different industry returns (dependent variable) and company-based combined sentiment. We used the distributed lag regression model on daily observations. We also report only the results of count-dependent sentiment indices made of NLTK scores. Log returns for the industries are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of each company and combining the returns in industry based indexes (market capitalization-weighted). The industries are: Health - healthcare (number of companies - 18), Fin - financial (15), Indust - industrials (14), IT - information technology (18), CD - consumer discretionary (12), CS - consumer staples (12) and Other (13). All regressions include constant, control variables of S&P 500 returns and its lags, as well as scaled daily mcap-weighted trading volumes of each company (Ind_TradingVolume) and its lags as well as lagged dependent variable. Independent variables are different for each regression and reflects the market-weighted combination of companies sentiment indices (Ind_SentIndex). Each independent variable has two lags. We corrected all regressions, except Reg_10, for the heteroscedasticity in residuals with heteroscedasticity consistent standard errors (HC3), because at least one of the tests rejected the homoscedasticity hypothesis (White test and Breusch–Pagan test). The stars indicate the significance of coefficients.

		Regressions with different industry dependent variables						
		Health	Fin	Indust	IT	CD	CS	Other
#	Independent variable	Reg_6	Reg_7	Reg_8	Reg_9	Reg_10	Reg_11	Reg_12
1	Ind_SentIndex	0.0030 *	0.0061 *	0.0026	-0.0037	0.0001	-0.0014	0.0000
	Ind_SentIndex_Lag_1	0.0002	-0.0008	-0.0002	0.0053 *	0.0033	-0.0018	-0.0029
	Ind_SentIndex_Lag_2	-0.0020	-0.0004	0.0016	-0.0059 **	-0.0032	0.0013	0.0006
2	S&P500returns	0.7184 ***	1.5946 ***	0.9819 ***	1.3348 ***	0.8140 ***	0.4686 ***	0.5957 ***
	S&P500returns_Lag_1	-0.0452	0.2184	0.0854	-0.1783	0.0071	-0.0404	-0.0218
	S&P500returns_Lag_2	-0.0829	-0.2646	0.1409	0.0396	0.0121	0.0340	0.0174
3	Ind_TradingVolume	-177.7159	63.6848	-39.1613	-202.4661	-67.9178	-32.2339	199.2198
	Ind_TradingVolume_Lag_1	189.5886	9.8957	-125.8496	78.6628	106.4310	87.5449	-31.0272
	Ind_TradingVolume_Lag_2	-225.6961	-3.1988	19.0320	21.8693	54.0879	190.5883	92.5650
4	DependentVar_Lag_1	0.0830	-0.0082	-0.0706	0.0536	-0.0106	-0.0168	0.1139
	DependentVar_Lag_2	0.1548 **	0.1726 **	-0.0647	-0.0144	0.0476	-0.0143	-0.0727
5	Constant	0.0004	-0.0059	-0.0030	0.0059	-0.0009	-0.0001	-0.0018
	# observations	250	250	250	250	250	250	250
	Adj. R-squared	0.3274	0.4760	0.5153	0.5599	0.1978	0.1319	0.2180

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

The aim of these regressions was to find whether Twitter sentiment related to the industry has an effect on the performance within the industry and which ones might be more sensitive. We can see that the industry sentiment indices (Ind_sentIndex) or their lags have a significant effect only for healthcare, financial and information technology at a 10% or 5% significance level. The other industries did not show significant sensitivity to sentiment. Concerning healthcare and financial, we can see that the contemporaneous effect of sentiment is positive for both industries and it might be an indication of short term reactions to the public opinion. As discussed in hypothesis 1, there are several reasons for why the contemporaneous variable might have a significant effect. The similar discussion could be used in this situation, too. If the contemporaneous variable has causal power embedded within, it is still hard to argue why these industries are more sensitive to the public opinion compared to consumer discretionary or consumer staple industries for example. One reasonable explanation could be that people tweet a lot more basic and irrelevant information about consumer industries and products. These tweets might neutralize the effect of the really significant tweets. The healthcare and financial industries might have a bigger share of more impactful tweets that can actually have some predictive power. In addition, these two industries had the lowest amount of tweets over the sample period (see Table 3). Following the discussion in section 5.1, it is possible that people who tweeted about these companies were more knowledgeable and thus their tweets carried more important information. Thus, since these industries possibly had a lower amount of irrelevant tweets, we were able to find significant coefficients. It is also apparent that, when using GICS industry specification, the companies attributed to the healthcare and financial industries are very similar to each other, mainly big pharma companies and financial institutions, respec-

tively. The same cannot be said in relation to the consumer industries as these groups are more diverse i.e. incorporate more different sub-industries. Therefore, a division in smaller subcategories for these industries could be interesting and might give more precise results. For example, literature found that aerospace is the industry that is most correlated with media exposure (W. Zhang, Skiena, et al., 2010). Thus as, the Aerospace players are part of Reg_8 (industrials) for which the coefficients do not show any significance, this might be an indication that we could have issues related to broad-categorization.

Furthermore, the Information technology (IT) industry has a significant lagged effect by its related sentiment index. The one trading day lag has a positive sign and the second has a negative. Both coefficients are significant at 10% and 5%, respectively. The signs might be and indication of lagging overreaction to sentiment. This industry also has the highest adjusted R-squared and by combining market performance and lagging sentiment, we can model almost 56% of the return variation within the industry. This finding is interesting in relation to other research findings and behavioral theory. Firstly, Baker and Wurgler (2006) found that companies, which are harder to value, are more sensitive to public sentiment. Even though, the companies in our sample are big corporations, the IT sector is known for being hard to predict regarding future cash flows, and thus harder to value. This result could therefore be somewhat consistent in relation to Baker and Wurgler's findings (2006). Secondly, Smailović et al. (2013) found that companies with higher return volatility also showed stronger predictability. As the IT industry is known for having volatile returns as well as it was the second most volatile industry in our sample (see Std for *Log returns: IT* in descriptive statistics, Table 3), we find support for Smailović et al. findings. Lastly, there is also signs that these results are in agreement with behavioral theory. For example the IT industry is generally more covered by media and usually more hyped (high salience). This can be supported by looking at the extraordinary high number of tweets within this industry [11] as well as the positivity of the tweets is among the highest (see table of descriptive statistics 3). According to one of the behavioral finance theories described in the literature overview, such hype should make the industry more prone for displaying signs of overreaction in prices. Thus, it could also explain why the IT industry had the best predictability. One also needs to note that the average daily return for the IT industry was by far the highest and could have influenced the results. Similarly, the results could possibly only be attributed to this year and may not be a finding that is consistent and reliable going forward.

Another important aspect to keep in mind, is that we used market capitalization-weighted indices in the industry analysis. In Table 11, the results of equally weighted combinations can be found. The signs of the coefficients are the same for all independent variables, except in the Other industry regression (Reg_12). We chose to focus on market-weighted results though, because it has additional underlying effect: the size of the companies might be important in the sentiment and return relationship modeling that is found by research (Brown & Cliff, 2004). Additional, we believe that sentiment of the bigger companies in the industry might have an effect on the industry itself, therefore a market weighted sentiment may be a smarter measure.

[11] This is also true if one takes into account market capitalization for the industries. Also, the size does not vary as much as the number of tweets.

5.4 Relationship between investor type and sentiment

To test the third hypothesis we grouped the company based sentiment and their corresponding stock returns, in a similar way as in industries part, only with a focus on retail investor share. We grouped the companies in our sample by taking 15 companies with the highest share of the retail investors (average share of retail investors among them - 38%) and 15 companies with the lowest (7%). The results can be seen in the Table 6. Contrary to our hypothesis, Reg_13 does not show any significant coefficients of the independent variable at 10% level. While the second lag of Reg_14 has a statistical significance at 1% level.

Table 6: H3: Regressions results based on Investor type

In this table we present the results of two regressions that try to model the relationship between different sentiment and returns interdependencies in relation to shareholders structure (investor type). We used distributed lag regression model on daily observations. We also report only the results of count-dependent sentiment indices made of NLTK scores. We divide regressions in two extremes: Reg_13 represents the relationship between sentiment and the 15 companies in our sample with the biggest share of retail investors, while Reg_14 represents the 15 companies with the biggest share of institutional investors and insiders (lowest retail investors share). Log returns of dependent variables are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of each company and combining the returns in indexes (market capitalization-weighted). All regressions include constant, control variables of S&P 500 returns and its lags, as well as scaled daily mcap-weighted trading volumes of each companies within the group (InvType_TradingVolume) and its lags as well as lagged dependent variable. Independent variables are different for each regression and reflects the market-weighted combination of companies sentiment indices (InvType_SentIndex). Each independent variable has two lags. We corrected both regressions for the heteroscedasticity in residuals with heteroscedasticity consistent standard errors (HC3), because at least one of the tests rejected the homoscedasticity hypothesis (White test and Breusch–Pagan test). The stars indicate the significance of coefficients.

		Regressions with different investor types	
		Top15Retail	Lowest15Retail
#	Independent variable	Reg_13	Reg_14
1	InvType_SentIndex	0.0031	0.0010
	InvType_SentIndex_Lag_1	0.0007	0.0003
	InvType_SentIndex_Lag_1	0.0020	-0.0027 ***
2	S&P500returns	0.6074 ***	1.1186 ***
	S&P500returns_Lag_1	-0.0035	0.0517
	S&P500returns_Lag_2	0.1330 *	0.0521
3	InvType_TradingVolume	34.2895	-863.5497
	InvType_TradingVolume_Lag_1	-141.5457	-109.3252
	InvType_TradingVolume_Lag_2	43.3319	-222.6324
4	DependentVar_Lag_1	0.0843	-0.0674
	DependentVar_Lag_2	-0.1390 *	-0.0434
5	Constant	-0.0046 *	0.0050 **
	# observations	250	250
	Adj. R-squared	0.3073	0.6744

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

The contemporaneous independent variables are insignificant at 10% level for both regressions. It is possible that concerning both of them, the prices are less efficient, and thus the overreaction as well as reversal to fundamentals have more lagging effect than compared to S&P 500 returns (discussed in H1). It is also important to note that different market participants have different trading motivations and patterns. Retail investors mainly think about the trading strategies after they come back home from the regular work, and thus

call brokers or place automated trades after trading hours (Antweiler & Frank, 2004). This might be a reason why there is no contemporaneous effect of sentiment in Reg_13. However, the signs of InvType_SentIndex coefficients are positive, which goes in lines with expectations.

In Table 6, we find that the second lag of the sentiment has a strong significant effect on the returns for companies with the highest share of institutional and insider investors. The other lags in both regressions are insignificant at 10% level. The significant lag has a negative coefficient, which might indicate the price correctness of previous overreaction. The found differences between these two regressions can be analyzed from several different points. First of all, this might be an indication that contrary to our previous assumptions, the companies with high degree of institutional investors can be easier predicted using sentiment. Thus, it is possible that we did not find predictive significance in Reg_13 because individual retail investors are less systematic in their investment decisions and InveType_SentIndex:Top15Retail variable incorporates many different investor motivations. The different type of incentives for building individual trading strategy might have diluted the general effect of sentiment. Second, as we have seen in descriptive statistics in Table 3, Top15Retail had a larger amount of tweets compared to Lowest15Retail variable. In relation to the discussion in section 5.1, it is possible that low amount of tweets might indicate that people who actually tweeted about those companies were more knowledgeable, and thus their sentiment had more impact on returns. The literature can add similar but, at the same time, a little bit different perspective. A study found that institutional investor-expressed sentiment had more predictive power compared to individual investor sentiment (Brown & Cliff, 2004). In addition, the researchers concluded that, contrary to the traditional opinion, institutional investor decisions might be easier affected by their own (institutional investor) expressed sentiment, compared to retail investors decisions. Therefore, by combining these two arguments in the literature, we could assume that institutional investor-expressed sentiment should have a stronger predictive power on the institutional investors decisions, and thus on the performance of companies with higher institutional investor base. Since we do not divide our Twitter data in institutionally expressed and individually expressed data, we can only guess what is the share of the institutional investors among Twitter participants in each group. However, we might argue that retail investors tweet less about the companies that they do not invest in, and thus InveType_SentIndex:Lowest15Retail had larger amount of tweets expressed by institutional investors. This would explain why lag of InveType_SentIndex:Lowest15Retail has higher significance compared to InveType_SentIndex:Top15Retail. Finally, the possibility exists that the difference in found significance between Reg_13 and Reg_14 are simply anomaly.

One side note in relation to companies size might be necessary here. The study found that lagged institutional sentiment had the strongest predictive power for large stocks (Brown & Cliff, 2004). Since we constructed the variables using market-weighted method, we indirectly incorporated the size of the companies in predicting the relationship (we gave bigger weight to larger companies, as it is in line with a literature). However, on average the size of companies with high retail investors share was larger as discussed in descriptive analysis. The better control for company size might be optimal in this case.

It is also interesting to see that the Adj. R-squared measure of Reg_14 explains two times more variations of the dependent variable compared to Reg_13. We have seen in Table 3 that Top15Retail average size of the company was double compared to Lowest15Retail. Thus, it is possible to assume that the bigger the company, the higher the complexity of the business model. Therefore, the prediction of its price movements might be more difficult. This reason might partly explain big differences in found Adj. R-squared measures between Reg_13 and Reg_14. As in the industry based regressions in the Table 5, the control variable of S&P 500 returns is a significant predictor of the relationship in the Table 6. Also, its economic significance is higher for Lowest15Retail compared to Tope15Retail. In addition, these regressions fail to find that the Trading Volume of those companies shares have a significant predictive power. Finally, it is important to note, that we assumed that retail investors share in companies did not change over time. This assumption might not hold in real life and thus, the results presented might be misleading.

5.5 Limitations of used models

In this part we want to shed a light on several general limitations of our models across all hypotheses. It is important to note that after adding sentiment and converting the data to the daily indices, we have only 250 observations in time series analysis. This still satisfies the rule of thumb of having the number of observations more than 10x independent variables in regression, however, it is not optimal. Thus, the results reported might have a lower robustness in this regard. Nevertheless, this is still a similar amount of observations compared to other established studies in this field (see Table 1) and our found results are consistent with their findings. Another limitation of our observations is the fact that it is daily. As discussed before, the probable finer grained effects are not visible in the reported results. In addition, one more limitation is the fact that we have only collected the data for the year 2017, which is known to have been a very positive year. The reported results might not be consistent in other years, especially in economic crises.

In addition, our measure of polarity might have a lower efficiency and hence lower significance of regression results. Bollen et al. (2011) found that sentiment division in mood dimensions (they used six dimensions) might be beneficial in constructing the indices. They found that especially 'calm' mood have a significant predictive power in explaining the returns of an asset. Therefore, by dividing our sentiment only in negative and positive mood, we might have neutralized certain sentiment effects. Moreover, the used sentiment analysis methods to classify sentiment of the text are not always accurate. Therefore, the misclassified sentiment scores might have led to somewhat spurious results in regressions. Similarly, by not using 'cashtag' in the collection of companies-related tweets, we might have gathered information that is not actually related to companies. This might add to the problem of spurious results. In addition, Bollen et al. (2011) found that using non-linear models increases the significance of the relationship between sentiment and market returns. Therefore, the linear regression might not be optimal in sentiment modeling and non-linear transformations might be required.

6 Conclusion and recommendations

With this paper we try to examine the effect of Twitter based sentiment on stock market returns. The aim was to add some new perspectives to the relatively young research field by using recently collected and expanded Twitter data as well as different methodologies. In addition to data and methodologies, we want to add some unique findings in explaining the general market returns, industry-based returns and ownership structure-based returns using sentiment.

Our main results show that sentiment, expressed over Twitter, can have additional value for explaining stock market returns. We find strong contemporaneous effect of different sentiment measures on S&P 500 returns, while the Bullishness measure shows stable significant predictive power. This is in line with found results in the literature for S&P 500. We also found that Vader NLTK sentiment analysis tool outperforms the traditionally used LM method. In addition, we find that by including different types of sentiment indices related to the general market, it adds value in explaining different characteristics of the S&P 500 price variations. However, our newly built company based index, which focuses more on non-financial information, has failed to add significant predictive information. After we divide the company-based sentiment in industries, we find relatively significant contemporaneous effect for the healthcare and financial industries. In addition, the sentiment of the IT industry displays significant predictive power. It seems to display the effect of both overreaction as well as reversal to fundamental value. Furthermore, contrary to the conventional assumption, companies with high share of retail investors fail to show higher sensitivity to sentiment. Instead, the sentiment measure indicates a statistically significant predictive ability for companies with the biggest share of institutional investors.

Several future recommendations in this field could be suggested, both from a data and methodology perspective as well as from the following research perspectives. From the data collection aspect, the time intervals could be more frequent than daily data, and the period of collected data could be longer as well, in order to investigate such relationship during both market expansions and contractions. In relation to the sentiment analysis, additional dimensions (not just positive and negative) could be considered. We also believe that the combination of Vader NLTK intensity and sentence characteristics techniques with the financial lexicon (LM) could add accuracy for financial sentiment analysis. In regards to the empirical methodology, the inclusion of tweet volumes could be considered, as well as a non-linear relationship model. Finally, further research could also investigate finer sub-industries as well as take a closer look at the relationship between firm size and sentiment.

References

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

Araujo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st annual acm symposium on applied computing* (pp. 1140–1145).

Aslam, S. (2018, 01 01). *Twitter by the numbers: Stats, demographics & fun facts.* `https://www.omnicoreagency.com/twitter-statistics/`. (Accessed: 2018-04-28)

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*(3), 307–343.

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

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

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

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017). 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)* (pp. 519–535).

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

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

EIKON. (2018). *Thomson Reuters EIKON.* [Online]. Available at: Subscription Service. (Accessed: 2018-03-16)

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

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

Gilbert, C., & Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media (icwsm-14).*

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

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

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

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

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

Lin, B., Zampetti, F., Bavota, G., Di Penta, M., Lanza, M., & Oliveto, R. (2018). Sentiment analysis for so ware engineering: How far can we go? *ICSE '18: 40th International Conference on Software Engineering , May 27-June 3, 2018, Gothenburg, Sweden..*

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

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65.

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

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, *29*(3), 305–325.

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

Mao, H., Counts, S., & Bollen, J. (2015). *Quantifying the effects of online bullishness on international financial markets* (Tech. Rep.). ECB Statistics Paper.

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

McGee, M. (2010, 03 31). *Twitter: How our new 'top tweets' works.* `https://searchengineland.com/twitter-how-our-new-top-tweets-works-39115`. (Accessed: 2018-04-15)

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

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

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

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

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

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 24–29).

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-computer interaction and knowledge discovery in complex, unstructured, big data* (pp. 77–88). Springer.

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

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014b). News or noise? using twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, *41*(7-8), 791–830.

Standard & Poor's. (2018). *Companies data.* Compustat database. (Accessed: 2018-03-16)

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, *8*(1), 113–116.

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

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

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Yahoo! Finance. (2018). *S&P 500 (GSPC) historic price data.* `https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC`. (Accessed: 2018-03-16)

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

Zhang, W., Skiena, S., et al. (2010). Trading strategies to exploit blog and news sentiment. In *Icwsm.*

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

Appendix

Appendix 1: Twitter and companies data table

Table 7: Twitter and companies data table

This table presents all the companies and other terms together with their corresponding tweets that were collected. Additionally, if applicable, we present the GICS Industry classification, retail investor concentration, search queries used and # of total tweets. We also show the percentage of positive tweets classified with the NLTK or LM sentiment analysis method for trading time. The last two columns exclude all the tweets that happened outside trading hours as well as neutral tweets. In total there are 102 companies, 3 other categories that all in all corresponds to approximately 95 million tweets collected for this study.

				Twitter data collection		Positivity of trading tweets	
#	Item	GICS Industry	Retail investors	Search query terms	# of Total Tweets	NLTK	LM
	Companies:						
1	Abbott Laboratories	Health Care	-	Abbott	733,907	52.11%	26.28%
2	AbbVie	Health Care	-	AbbVie	19,549	73.44%	54.38%
3	Accenture	Information Technology	-	Accenture	142,444	87.50%	72.02%
4	Activision Blizzard	Information Technology	Low	Activision&Blizzard	15,279	70.33%	52.97%
5	Adobe Systems	Information Technology	Low	Adobe	704,913	77.55%	70.19%
6	Aetna	Health Care	-	Aetna	12,078	61.25%	36.32%
7	Allergan	Health Care	-	Allergan	21,875	69.31%	40.66%
8	Altria Group	Consumer Staples	High	Altria	15,580	71.97%	52.35%
9	Amazon.com	Consumer Discretionary	-	Amazon	14,426,477	83.16%	55.87%
10	American Express	Financials	-	American&Express	139,994	71.02%	50.13%
11	American International Gr	Financials	-	American&International, AIG	27,451	73.70%	47.97%
12	American Tower	Real Estate	Low	American&Tower	3,762	81.65%	70.61%
13	Amgen	Health Care	-	Amgen	28,246	69.04%	48.77%
14	Applied Materials	Information Technology	-	Applied&Materials	5,571	84.26%	62.56%
15	AT&T	Telecommunication Services	High	AT&T	1,331,303	62.99%	36.21%
16	Automatic Data Processing	Information Technology	-	Automatic&Data&Processing, ADP	151,670	77.69%	57.94%
17	Bank of America	Financials	-	Bank&of&America, BofA	396,200	64.78%	41.33%
18	Bank of New York Mellon	Financials	-	BNY&Mellon	5,796	86.77%	69.95%
19	Becton, Dickinson and Co	Health Care	-	Becton&Dickinson	5,166	79.88%	67.31%
20	Berkshire Hathaway	Financials	High	Berkshire&Hathaway	36,077	82.53%	59.65%
21	Biogen	Health Care	Low	Biogen	7,698	83.02%	60.59%
22	BlackRock	Financials	Low	BlackRock	92,513	71.86%	49.07%
23	Boeing	Industrials	-	Boeing	578,261	65.16%	31.90%
24	Bristol-Myers Squibb	Health Care	-	Bristol&Myers, Bristol-Myers	13,974	63.12%	50.78%
25	Broadcom	Information Technology	Low	Broadcom	43,928	49.39%	20.90%
26	Celgene	Health Care	-	Celgene	18,345	66.34%	53.95%
27	Charles Schwab	Financials	Low	Charles&Schwab	19,739	79.67%	63.50%
28	Charter Communications	Consumer Discretionary	Low	Charter&Communications	6,838	74.20%	42.32%
29	Chevron	Energy	High	Chevron	178,759	72.71%	41.97%
30	Cigna	Health Care	Low	Cigna	42,500	58.36%	36.89%
31	Cisco Systems	Information Technology	-	Cisco	746,333	78.37%	57.10%
32	Citigroup	Financials	-	Citigroup, Citi, Citibank	326,789	68.08%	43.56%
33	Coca-Cola	Consumer Staples	High	CocaCola, Coca&Cola, Coca-Cola	599,944	71.82%	50.93%
34	Colgate-Palmolive	Consumer Staples	-	Colgate-Palmolive, Palmolive, ColgatePalmolive, Colgate	44,032	78.07%	56.78%
35	Comcast	Consumer Discretionary	-	Comcast	733,845	56.05%	27.83%
36	ConocoPhillips	Energy	-	ConocoPhillips, Conoco&Phillips	8,335	75.47%	40.93%
37	Costco Wholesale	Consumer Staples	-	Costco	581,786	68.13%	44.63%
38	CSX	Industrials	-	CSX	47,843	59.16%	32.05%
39	CVS Health	Consumer Staples	-	CVSHealth, CVS&Health, CVS&Corporation	32,511	78.62%	53.57%
40	Deere	Industrials	-	Deere	88,113	75.74%	60.07%
41	Duke Energy	Utilities	High	Duke&Energy	91,528	74.83%	39.35%
42	Eli Lilly	Health Care	-	Eli&Lilly	14,988	57.40%	37.23%
43	Exxon Mobil	Energy	High	Exxon&Mobil, Exxon, ExxonMobil	795,286	59.01%	33.22%
44	Facebook	Information Technology	-	Facebook	20,531,802	66.52%	43.50%
45	FedEx	Industrials	-	FedEx	317,111	59.49%	32.24%
46	Ford Motor	Consumer Discretionary	High	Ford&Motor, FordMotor, Ford&Motor	9,667	75.70%	48.55%
47	General Dynamics	Industrials	Low	General&Dynamics	8,351	91.71%	47.47%
48	General Electric	Industrials	High	General&Electric	87,904	68.06%	50.12%
49	General Motors	Consumer Discretionary	-	General&Motors	9,556	65.98%	37.81%
50	Gilead Sciences	Health Care	-	Gilead&Sciences, Giliead	65,693	63.79%	40.57%
51	Goldman Sachs Group	Financials	-	Goldman&Sachs	307,738	57.30%	29.73%
52	Home Depot	Consumer Discretionary	-	Home&Depot	433,864	66.54%	46.71%
53	Honeywell International	Industrials	-	Honeywell	100,492	83.63%	59.52%
54	IBM	Information Technology	High	IBM	1,069,582	80.94%	60.52%
55	Illinois Tool Works	Industrials	-	Illinois&Tool&Works, ITW	13,892	84.57%	57.19%
56	Intel	Information Technology	-	Intel	2,218,294	52.49%	22.86%
57	Johnson & Johnson	Health Care	-	Johnson & Johnson, Johnson&Johnson	62,347	63.23%	43.08%
58	JPMorgan Chase	Financials	-	J.P.Morgan, JPMorgan, JP&Morgan, J.P.&Morgan	242,667	65.53%	32.95%
59	Kraft Heinz	Consumer Staples	-	KraftHeinz, Kraft&Heinz	20,738	55.17%	28.71%
60	Lockheed Martin	Industrials	-	Lockheed&Martin	103,735	71.20%	43.36%
61	Lowe's Companies	Consumer Discretionary	-	Lowe&s Companies	4,873	77.66%	64.77%
62	Mastercard	Information Technology	Low	Mastercard	184,613	77.28%	61.09%
63	McDonald's	Consumer Discretionary	-	McDonald%s	2,850,585	62.71%	40.56%
64	Medtronic	Health Care	-	Medtronic	27,353	76.94%	54.01%
65	Merck & Co	Health Care	-	Merck	118,954	45.90%	26.93%
66	Microsoft	Information Technology	-	Microsoft	3,743,010	71.50%	50.60%
67	Mondelez International	Consumer Staples	-	Mondelez	6,455	69.16%	47.85%
68	Monsanto	Materials	-	Monsanto	218,906	36.20%	13.71%
69	Morgan Stanley	Financials	-	Morgan&Stanley	133,785	64.51%	40.44%
70	Netflix	Information Technology	-	Netflix	10,444,487	64.50%	40.73%
71	NextEra Energy	Utilities	-	NextEra	5,564	84.45%	38.84%
72	Nike	Consumer Discretionary	-	Nike	3,474,930	69.22%	44.47%
73	Northrop Grumman	Industrials	-	Northrop&Grumman	43,836	79.84%	49.48%
74	NVIDIA	Information Technology	-	NVIDIA	379,449	76.53%	58.09%
75	Occidental Petroleum	Energy	-	Occidental&Petroleum	4,345	81.38%	58.27%
76	PayPal Holdings	Information Technology	-	PayPal	1,580,622	71.15%	60.12%
77	PepsiCo	Consumer Staples	-	PepsiCo, Pepsi	1,087,004	63.23%	40.79%
78	Pfizer	Health Care	-	Pfizer	75,287	65.82%	54.16%
79	Philip Morris Intl	Consumer Staples	-	Philip&Morris	16,085	62.96%	37.89%

(continued on the next page)

(continued)

#	Item	GICS Industry	Retail investors	Twitter data collection: Search query terms	Twitter data collection: # of Total Tweets	Positivity of trading tweets: NLTK	Positivity of trading tweets: LM
	Companies:						
80	Phillips 66	Energy	-	Phillips&66, Phillips66	8,936	68.91%	42.47%
81	PNC Financial Services Gr	Financials	-	PNC&Financial	15,516	72.54%	56.95%
82	Procter & Gamble	Consumer Staples	High	Procter&Gamble, Procter & Gamble, P&G	2,574,189	68.12%	45.56%
83	Prudential Financial	Financials	High	Prudential&Financial	7,270	85.61%	73.92%
84	Qualcomm	Information Technology	-	Qualcomm	211,377	55.00%	30.92%
85	Raytheon	Industrials	-	Raytheon	57,883	69.87%	36.94%
86	Salesforce.com	Information Technology	Low	Salesforce.com, Salesforce	402,797	87.44%	72.51%
87	Schlumberger	Energy	-	Schlumberger	13,886	70.57%	42.27%
88	Simon Property Group	Real Estate	Low	Simon&Property	5,770	76.91%	55.76%
89	Starbucks	Consumer Discretionary	-	Starbucks	3,833,938	67.88%	44.78%
90	Texas Instruments	Information Technology	-	Texas&Instruments	20,347	82.53%	68.69%
91	Thermo Fisher Scientific	Health Care	Low	Thermo&Fisher	16,526	84.66%	69.74%
92	Time Warner	Consumer Discretionary	-	Time&Warner, TimeWarner	39,721	50.67%	18.88%
93	Union Pacific	Industrials	-	Union&Pacific	21,466	72.53%	47.69%
94	United Parcel Service	Industrials	High	United&Parcel&Service, UPS	2,211,655	64.56%	43.17%
95	United Technologies	Industrials	-	United&Technologies	8,835	93.00%	45.33%
96	UnitedHealth Group	Health Care	Low	UnitedHealth	19,233	71.62%	45.86%
97	US Bancorp	Financials	-	U.S.&Bank, US&Bancorp, US&Bank	77,745	69.91%	45.59%
98	Verizon Communications	Telecommunication Services	High	Verizon	1,114,743	59.14%	33.50%
99	Wal-Mart Stores	Consumer Staples	-	Wal-Mart, WalMart	2,403,566	61.49%	40.65%
100	Walgreens Boots Alliance	Consumer Staples	-	Walgreens	331,926	65.55%	48.67%
101	Walt Disney	Consumer Discretionary	High	WaltDisney, Disney	3,467,887	74.59%	62.82%
102	Wells Fargo	Financials	-	Wells&Fargo	406,990	67.77%	22.57%
	Other:						
1	S&P 500	-	-	s&p&500, s&p500, sp&500, sp500, Standar&Poor&500, the&S&P	189,389	63.01%	33.90%
2	Bullishness*	-	-	bearish, bear&market, bullish, bull&market	704,509	-	-
3	Market terms	-	-	equities, indexes, stock&market, stock&markets, stocks	3,429,457	62.43%	36.47%
	Average				**904,156**	**70.24%**	**46.88%**
	Total number of tweets				**94,936,359**	**15,343,900****	**7,296,360****

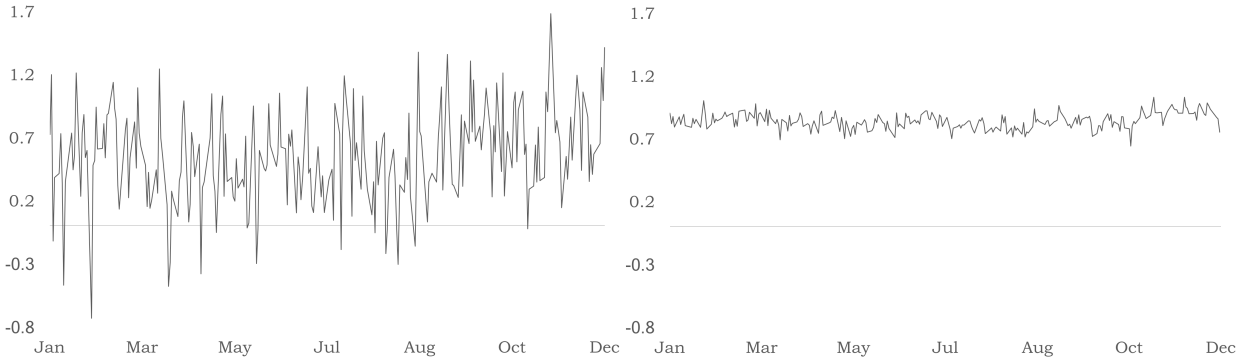
*Positivity = 73.75%

**Excluding Bullishness tweets

Appendix 2: Sentiment indices for H1

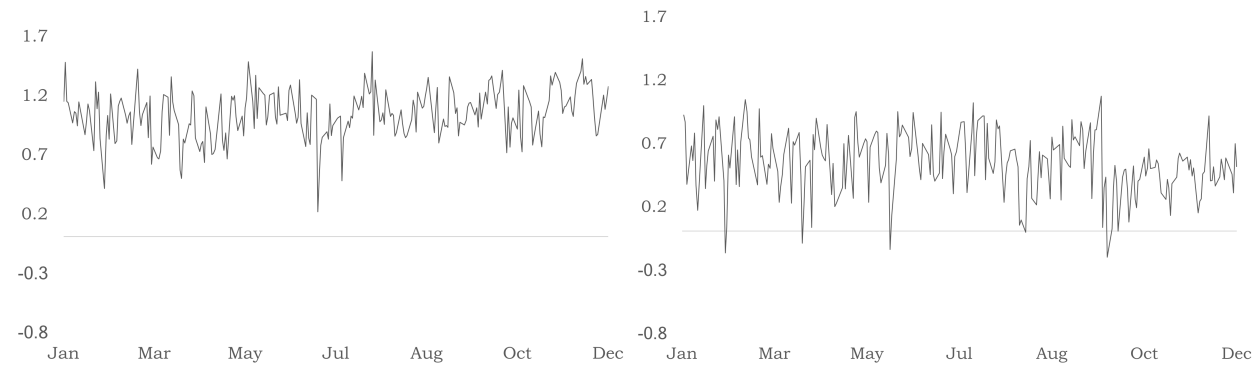
Figure 3: Sentiment indices for our first hypothesis

In this Figure the four main (S& 500, Bullishness, Market Terms, Companies) sentiment indices are presented. The aim of this figure is too visualize the four different indices and especially their differences. It shows count-dependent NLTK indices.



(a) S&P500 sentiment index

(b) Companies sentiment index



(c) Bullishness sentiment index

(d) Market Terms sentiment index

Appendix 3: Results of other regressions for H1

Table 8: H1: Regressions results with count-neutral NLTK

In this table we present the results of five different regressions that try to model the relationship between dependent variable (S&P 500) log returns and selected independent variables. We used distributed lag regression model on daily observations. Log returns are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of S&P 500. All regressions include constant, control variables of scaled daily S&P 500 Trading Volume and its lags as well as lagged dependent variable (S&P 500 log returns). Independent variables - NLTK_S&P500 and NLTK_MarketTerms - are count-neutral sentiment indices made of NLTK scores. NLTK_S&P500 uses the tweets data that was collected using S&P 500-related search queries and NLTK_MarketTerms variable refers to collected data of general financial market terms (the list of words can be found in Appendix). Bullishness is count-neutral index of 'bullish' and 'bearish' Twitter data (no additionally added sentiment). NLTK_CompaniesIndex is the market capitalization-weighted index comprised of each of the 102 biggest S&P companies count-neutral sentiment indices of NLTK scores. Each independent variable has two lags. The stars indicate the significance of coefficients.

		Regressions with dependent variable - S&P 500 log returns				
#	Independent variable	Reg_1	Reg_2	Reg_3	Reg_4	Reg_5
1	NLTK_S&P500	0.0129 ***				0.0059 ***
	NLTK_S&P500_Lag_1	-0.0009				0.0006
	NLTK_S&P500_Lag_2	0.0009				0.0011
2	Bullishness		0.0279 ***			0.0163 ***
	Bullishness_Lag_1		-0.0097 ***			-0.0071 **
	Bullishness_Lag_2		-0.0016			0.0019
3	NLTK_MarketTerms			0.0211 ***		0.0132 ***
	NLTK_MarketTerms_Lag_1			-0.0055 **		-0.0023
	NLTK_MarketTerms_Lag_2			-0.0022		-0.0011
4	NLTK_CompaniesIndex				0.0155 *	
	NLTK_CompaniesIndex_Lag_1				-0.0068	
	NLTK_CompaniesIndex_Lag_2				0.0103	
5	TradingVolume	0.0666	-0.1107	0.0067	-0.0119	0.0379
	TradingVolume_Lag_1	1.2188 **	1.1753 **	0.8282 *	1.0512 *	0.9867 **
	TradingVolume_Lag_2	-0.9542 **	-0.3917	-1.0398 **	-1.5233 ***	-0.5223
6	S&P500returns_Lag_1	-0.2026 ***	-0.1097 *	-0.1545 **	-0.1257 **	-0.1851 ***
	S&P500returns_Lag_2	-0.1215 *	-0.0391	-0.0428	-0.0666	-0.1206 **
7	Constant	-0.0002	-0.0094 ***	-0.0020	-0.0053	-0.0086 ***
	# observations	250	250	250	250	250
	Adj. R-squared	0.2907	0.2921	0.3264	0.0431	0.4820

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

Table 9: H1: Regressions results with count-dependent LM

In this table we present the results of five different regressions that try to model the relationship between dependent variable (S&P 500) log returns and selected independent variables. We used distributed lag regression model on daily observations. Log returns are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of S&P 500. All regressions include constant, control variables of scaled daily S&P 500 Trading Volume and its lags as well as lagged dependent variable (S&P 500 log returns). Independent variables - cLM_S&P500 and cLM_MarketTerms - are count-dependent sentiment indices made of LM scores. cLM_S&P500 uses the tweets data that was collected using S&P 500-related search queries and cLM_MarketTerms variable refers to collected data of general financial market terms (the list of words can be found in Appendix). cBullishness is count-dependent index of 'bullish' and 'bearish' Twitter data (no additionally added sentiment). cLM_CompaniesIndex is the market capitalization-weighted index comprised of each of the 102 biggest S&P companies count-dependent sentiment indices of LM scores. Each independent variable has two lags. The stars indicate the significance of coefficients.

		Regressions with dependent variable - S&P 500 log returns				
#	Independent variable	Reg_1	Reg_2	Reg_3	Reg_4	Reg_5
1	cLM_S&P500	0.0033 ***				0.0015 ***
	cLM_S&P500_Lag_1	-0.0002				-0.0009 *
	cLM_S&P500_Lag_2	-0.0004				0.0004
2	cBullishness		0.0110 ***			0.0091 ***
	cBullishness_Lag_1		-0.0037 ***			-0.0039 ***
	cBullishness_Lag_2		-0.0010			0.0004
3	cLM_MarketTerms			0.0074 ***		0.0053 ***
	cLM_MarketTerms_Lag_1			-0.0013		-0.0005
	cLM_MarketTerms_Lag_2			-0.0011		-0.0007
4	cLM_CompaniesIndex				-0.0032	
	cLM_CompaniesIndex_Lag_1				0.0036	
	cLM_CompaniesIndex_Lag_2				0.0049	
5	TradingVolume	-0.0140	-0.1169	-0.0659	-0.1996	-0.1148
	TradingVolume_Lag_1	0.8966 *	1.1749 **	0.8318 *	1.3293 **	0.8333 *
	TradingVolume_Lag_2	-1.0843 **	-0.3950	-1.1865 ***	-1.2873 **	-0.4498
6	S&P500returns_Lag_1	-0.1553 **	-0.1165 *	-0.1545 **	-0.1356 **	-0.1084 *
	S&P500returns_Lag_2	-0.0821	-0.0311	-0.0680	-0.0527	-0.0898
7	Constant	0.0015	-0.0080 ***	0.0049 **	0.0020	-0.0038
	# observations	250	250	250	250	250
	Adj. R-squared	0.1214	0.2871	0.2244	0.0341	0.4204

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

Table 10: H1: Regressions results with count-neutral LM

In this table we present the results of five different regressions that try to model the relationship between dependent variable (S&P 500) log returns and selected independent variables. We used distributed lag regression model on daily observations. Log returns are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of S&P 500. All regressions include constant, control variables of scaled daily S&P 500 Trading Volume and its lags as well as lagged dependent variable (S&P 500 log returns). Independent variables - LM_S&P500 and LM_MarketTerms - are count-neutral sentiment indices made of LM scores. LM_S&P500 uses the tweets data that was collected using S&P 500-related search queries and LM_MarketTerms variable refers to collected data of general financial market terms (the list of words can be found in Appendix). Bullishness is count-neutral index of 'bullish' and 'bearish' Twitter data (no additionally added sentiment). LM_CompaniesIndex is the market capitalization-weighted index comprised of each of the 102 biggest S&P companies count-neutral sentiment indices of LM scores. Each independent variable has two lags. The stars indicate the significance of coefficients.

		Regressions with dependent variable - S&P 500 log returns				
#	Independent variable	Reg_1	Reg_2	Reg_3	Reg_4	Reg_5
1	LM_S&P500	0.0068 ***				0.0030 **
	LM_S&P500_Lag_1	-0.0004				-0.0020
	LM_S&P500_Lag_2	-0.0011				0.0008
2	Bullishness		0.0279 ***			0.0230 ***
	Bullishness_Lag_1		-0.0097 ***			-0.0099 ***
	Bullishness_Lag_2		-0.0016			0.0018
3	LM_MarketTerms			0.0170 ***		0.0123 ***
	LM_MarketTerms_Lag_1			-0.0028		-0.0010
	LM_MarketTerms_Lag_2			-0.0025		-0.0017
4	LM_CompaniesIndex				-0.0021	
	LM_CompaniesIndex_Lag_1				0.0048	
	LM_CompaniesIndex_Lag_2				0.0103	
5	TradingVolume	0.0647	-0.1107	-0.0807	-0.1673	-0.1159
	TradingVolume_Lag_1	0.9776 *	1.1753 **	0.8264 *	1.2386 **	0.8702 **
	TradingVolume_Lag_2	-1.1840 **	-0.3917	-1.1786 ***	-1.2407 **	-0.4982
6	S&P500returns_Lag_1	-0.1523 **	-0.1097 *	-0.1587 **	-0.1346 **	-0.1115 *
	S&P500returns_Lag_2	-0.0762	-0.0391	-0.0695	-0.0569	-0.0935
7	Constant	0.0029	-0.0094 ***	0.0053 **	0.0020	-0.0047 *
	# observations	250	250	250	250	250
	Adj. R-squared	0.1070	0.2921	0.2365	0.0344	0.4245

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level

Appendix 4: Equally-weighted industry regression results for H2

Table 11: H2: Equally-weighted Industry Regressions results

In this table we present the results of seven regressions that try to model the relationship between different industries returns (dependent variable) and company-based combined sentiment. We used distributed lag regression model on daily observations. We also report only the results of count-dependent sentiment indices made of NLTK scores. Log returns of industries are calculated by taking the natural logarithm of the quotient of the consecutive closing prices of each company and combining the returns in industry based indexes (equally-weighted). The industries are: Health - health-care (number of companies - 18), Fin - financial (15), Indust - industrials (14), IT - information technology (18), CD - consumer discretionary (12), CS - consumer staples (12) and Other (13). All regressions include constant, control variables of S&P 500 returns and its lags, as well as scaled daily equally-weighted trading volumes of each company (Ind_TradingVolume) and its lags as well as lagged dependent variable. Independent variables are different for each regression and reflects the equally-weighted combination of companies sentiment indices (Ind_SentIndex). Each independent variable has two lags. We corrected all regressions, except Reg_10 and Reg_12, for the heteroscedasticity in residuals with heteroscedasticity consistent standard errors (HC3), because at least one of the tests rejected the homoscedasticity hypothesis (White test and Breusch–Pagan test). The stars indicate the significance of coefficients.

		Regressions with different industry dependent variables						
		eqHealth	eqFin	eqIndust	eqIT	eqCD	eqCS	eqOther
#	Independent variable	Reg_6	Reg_7	Reg_8	Reg_9	Reg_10	Reg_11	Reg_12
1	eqInd_SentIndex	0.0030 *	0.0020	0.0017	-0.0021	0.0021	-0.0014	-0.0001
	eqInd_SentIndex_Lag_1	0.0003	-0.0007	-0.0006	0.0023	0.0023	-0.0007	-0.0020
	eqInd_SentIndex_Lag_2	-0.0028	0.0027	0.0007	-0.0029	-0.0022	0.0001	-0.0001
2	S&P500returns	0.7616 ***	1.5049 ***	1.0151 ***	1.3861 ***	0.7684 ***	0.5231 ***	0.5512 ***
	S&P500returns_Lag_1	-0.0712	0.2546 **	0.0327	-0.2021	0.0946	-0.0147	0.0106
	S&P500returns_Lag_2	-0.0850	-0.2715 *	0.1649	-0.0068	0.0637	0.0525	-0.0549
3	eqInd_TradingVolume	-359.5846	108.9620	178.9130	-608.1752	-36.8521	-206.8678	386.5065
	eqInd_TradingVolume_Lag_1	322.6691	24.4007	-288.8217	201.0244	4.1655	206.6055	24.2756
	eqInd_TradingVolume_Lag_2	-332.2687	-41.3595	61.4569	38.6877	54.6553	173.7563	-46.2549
4	eqDependentVar_Lag_1	0.0974	-0.0537	-0.0370	0.0254	-0.0949	-0.0040	0.0231
	eqDependentVar_Lag_2	0.1468 **	0.1990 ***	-0.0802	0.0030	0.0366	-0.0316	-0.0209
5	Constant	0.0016	-0.0052	-0.0014	0.0067 *	-0.0019	0.0004	-0.0012
	# observations	250	250	250	250	250	250	250
	Adj. R-squared	0.3431	0.5242	0.5082	0.5646	0.2584	0.1460	0.2210

* Indicate Significance at the 10% level

** Indicate Significance at the 5% level

*** Indicate Significance at the 1% level