LUND
UNIVERSITY

School of Economics and Management

NEKN02: Master Essay

Students:

Levi Bergstrand

Maksim Khritantsev

Supervisor:

Olga Balakina

Lund, 2018

How 140 Characters can be related to the Stock Market Movements:

Sentiment Analysis of Twitter

Levi Bergstrand **(93-04-07)**

Maksim Khritantsev **(95-10-25)**

A Thesis in the Field of Finance

For the Degree of Master of Science in Finance

Lund University

June 2018

Abstract

Stock market movements forecast based on sentiment analysis is certainly a field worth investigating. Being able to build future investment strategies based on forecasted stock returns would be of tremendous importance for individual investors and high-frequency trading firms. This thesis aims to closely investigate the impact that tweet sentiments in Twitter have on stock market movements; it intends to empirically test the prediction power of public sentiments of tweets with respect to stock price returns of S&P 500 companies. In the analysis of the study, this paper uses a sample consisting of 181 776 tweets, collected between December 26th, 2017 and March 15th, 2018. The results of the dissertation present significant evidence of dependence between tweet sentiments and stock returns, using lexicon-based panel data regression. The closing results show, using complex algorithmic machine learning techniques, that random forest and SVM-Gaussian are the optimal models in the prediction of stock returns based on tweets under unigram, bigram and trigram methods.

Keywords

stock returns;
tweet sentiments;
lexicon based; panel data;
regression analysis; machine learning;
support vector machines;
random forest;

Table of Contents

Tables - Figures - Graphs

1. Introduction

Social networks are present in numerous spheres of people's lives, thus, they have an enormous capacity to alter the approach towards socio-economic sciences. Twitter is arguably one of the most powerful means of sharing information on the Internet. This micro-blogging platform has had 330 million monthly active users by the 4th quarter of 2017, which makes it the 12th most visited website in the world, according to Alexa ranking.

Twitter offers all of its users the opportunity to express opinions, ideas, judgments, concerns, and attitudes regarding any kind of cultural, political or economic event. This is featured in the work of Tumasjan, Sprenger, Sandner, and Welpe (2010) who discover that the number of times a particular party is mentioned on Twitter, has a clear relation with the outcome of the German federal elections. In the financial sphere, Bollen, Mao, and Pepe (2011) prove that positive or negative sentiments derived from a large-scale collection of tweets serving as a reliable measure of the public's mood are reflected in the Dow Jones Industrial Average with a lag ranging from two to six days.

Therefore, Twitter does have a predictive power with respect to various socio-economic events, including the shifts in the stock markets. In fact, we have already witnessed cases such as the hack of the Associated Press Twitter post about President Obama's injury resulting in $136 billion (USD) loss of S&P 500 value, and Donald Trump's post on Lockheed Martin leading to a 4% loss of the company's stock value. That is why the primary goal of this thesis is to empirically test if the overall public sentiment derived from the tweets, can be a good way to forecast the stock returns of companies included in the S&P 500 index. This will be accomplished by applying various techniques; starting with the Opinion-Finder, which measures positive versus negative sentiments, and moving on to more advanced Machine Learning techniques.

The main contribution of this dissertation is to measure the public sentiment reflected by Twitter and to test its effect on the stock market. This is done by pooling an extensive sample of Twitter posts related to stocks we are interested in, and by utilizing the predictive capacity of widely used methods such as the Sentiment Index calculation and other techniques previously underestimated in analyzing Twitter's relation with stock return time-series.

This paper also holds a practical importance for individual investors and high-frequency trading firms. If Twitter is capable of accumulating valuable information, which reflects public moods and opinions, it can be used to build future investment strategies around predicted stock returns.

The dissertation is divided into the following parts: chapter 2 addresses existing papers devoted to this topic. Chapter 3 touches upon data collection and cleaning and presents data overview. Chapter 4 covers the methodology of the paper. Chapter 5 lists the results of the thesis. Chapter 6 provides the conclusion of the study, discusses the limitations and presents ideas for future research.

2. Literature review

The arguments about possibilities of predicting stock prices are dating back to the first mentioning of the efficient market hypothesis (EMH) concept. The EMH is an investment theory in financial economics that states that it is impossible to outperform the market since stocks are already perfectly priced and reflect all available information. This implicates that it is theoretically impossible to achieve gains from any trading strategy (arbitrage, swing trading, trend lines, candlestick patterns, etc.). Consequently, since there is no way to rely on fundamental or technical analysis to predict future prices, purchasing higher risk investments is the only way to earn higher returns than those of an index (Fama, 1965).

First and foremost, we are going to present some papers about the efficient market hypothesis uncertainty and its reflection in economic and financial phenomena. On this topic, we are going to focus on the two following articles: 'A Non-Random Walk Down Wall Street' and 'Foundations of Technical Analysis'.

There have been numerous studies attempting to question the EMH. Lo and MacKinlay (1999) empirically find that short-term serial correlations in security time-series, i.e. consistent price movements in one direction, do not go in line with the random walk hypothesis, thus, rejecting the EMH. Later on, this discovery is proved by Lo, Mamaysky, and Wang (2000) who claim that investors using complicated technical analysis methods to detect patterns in stock price movements, can also help in gaining the momentum (above average stock returns). The authors apply bootstrapping and Monte Carlo methods to several hundred U.S. stocks from 1962 to 1996. According to them, statistical techniques like 'head and shoulders' (Lo, Mamaysky, and Wang, 2000 – p. 11) and 'double bottoms' (Lo, Mamaysky, and Wang, 2000 – p. 12) are likely to be predictable of stock prices.

Nonetheless, the findings of behavioral economists about irrational human behavior act as another contradiction to the EMH. Institutional investors are often prone to the gambler's fallacy. They perceive the future stock returns as inversely proportional to the current returns; investors might underestimate the probability of having high returns in the future, in case they have already observed relatively high returns in the past (Shefrin, 2007). According to Malkiel, Mullainathan, and Stangle (2005), the asset bubble emerges when market participants are not willing to sell the mispriced assets due to losses they have already endured, which is an example of a sunk-cost fallacy. Ultimately, irrational behavior embedded in investment decisions prevents individuals and firms from making correct judgments over the given assets.

Thereupon, many papers are devoted to the topic of efficient market hypothesis uncertainty that finds reflection in economic and financial phenomena. Further, in our paper, we aim to disprove the EMH by showing that stock returns can be predicted by using information about public sentiment from Twitter.

We are also going to elaborate on how public sentiment finds a reflection in political and economic phenomena. On this topic, we will be focusing on the two following articles: 'The Impact of Political Uncertainty on Asset Prices' and 'Predicting the Future with Social Media'.

There have been different studies about the effect that political uncertainty has on asset pricing, and how any political instability is impacted by people's emotional state on the stock market. It is demonstrated in theoretical models how risk premium is demanded due to political uncertainty, and how the magnitude of the risk premium is larger in fragile economic conditions (Pastor and Veronesi, 2013). Researchers have tested various different empirical techniques to create a political uncertainty index. Pastor and Veronesi use a political uncertainty index to validate a political risk premium estimated by their model. Lui, Shu, and Wei (2014) trace a natural experiment and use the political uncertainty index as evidence to validate their study. The shocking political incident of Bo's scandal of China, 2012, and the impact of such political uncertainty on asset pricing, is the main result of their paper. The former Communist Party chief of Chongqing China, Bo Xilai, who was once perceived as a top office runner, was facing prosecution along with his wife who was suspended a death sentence for the murder of a British executive. Bo's event was very significant for the stability of the country, especially that there was a lot of uncertainty about whether the transference of authority would be peaceful or not. According to the equilibrium model built by Pastor and Veronesi (2013), stock prices are expected to drop with announcements of policy changes; the higher the uncertainty about a governmental policy change, the higher the decrease in the stock price. This scandal case, in particular, delivers a perfect workshop setting to test the fundamental link between political uncertainty and asset prices, for there is a positive date when the political uncertainty suddenly increases. The political uncertainty risk model anticipates that rises in political uncertainty, causes stock prices to decrease.

Moreover, our thesis addresses the EMH uncertainty in an attempt to point out that the EMH does not admit that the public mood can be predictive of security prices. Nevertheless, people's emotions are proved to have relations with financial indicators (Nofsinger, 2005).

In his paper, Nofsinger states that optimistic/pessimistic attitudes are reflected in consumers' and investors' decisions, and besides that, the tone of business activities is influenced by the public mood, but not vice versa.

Hence, information from social networks, which are the encapsulated versions of public sentiments, can be used to track and predict stock returns. One of the first papers devoted to the topic of using tweets for quantitative research is Asur and Huberman (2010), in which the authors use the content of Twitter to forecast movie revenues. The authors estimated a linear regression model on the rate of chatter from three million tweets from Twitter. The main result of the paper is that the attention and popularity of films derived from Twitter's positive/negative sentiments are far better predictors of revenue streams than the Hollywood Stock Exchange index.

Upon that, various research papers are dedicated to the subject matter of public sentiment that finds reflection in political and economic phenomena. Yet, in our paper, we intend to provide evidence that public sentiment can be related to such a financial phenomenon as stock returns change.

The third and final topic we are elaborating on in literature review is about the attention that the relation between social networks and socio-economic indicators has been receiving. On this subject, we study the six following articles, respectively: 'Twitter Mood Predicts the Stock Market', 'The Effect of Twitter Sentiment on Stock Price Returns', 'Predicting Stock Market Indicators through Twitter', 'Predicting the Present with Google Trends', 'Widespread Worry and the Stock Market' and 'Social Media Sentiment Analysis using Machine Learning Classifiers'.

Firstly, we thoroughly investigate the work of Bollen J., Mao H. and Zeng X. (2010) who explore the relation between tweets and the DJIA index. In their study, the authors collect ten million tweets from the year 2008 and categorize them according to the Google profile of Mood States: calmness, anxiety, confidence, energy, kindness, and happiness. After running several time-series regressions, it turns out that the 'calm' time-series can predict the DJIA index lagged by +3 days with the accuracy of 88%. Another method they implement is to divide tweets into positive and negative, and then to apply the Granger Causality test with lags in the range of two to six days. The results prove again that tweet sentiments can be used to forecast the DJIA index. Our paper will include a slightly modified version of the binary sentiment classification used by Bollen et. al. (2010). That is why this paper is highly relevant to our thesis.

We also look into Ranco, Aleksovski, Caldarelli, Grear and Mozeti (2015) who as Bollen et. al. (2010), apply the Granger Causality test to validate the nature of the established relation. In addition, they use the Pearson correlation and event studies to further investigate the issue. This study is performed using supervised machine learning and bag-of-word techniques in order to evaluate the sentiments contained in 1.5 million tweets. Even though the authors do not find a strong reliance between stock returns and Twitter sentiments, they discover a significant correlation, persisting over 15 consecutive months, between cumulative abnormal returns and the volume of tweets mentioning selected DJIA companies.

Moving further, we study how Zhang, Fueres, and Gloor (2010) take a slightly different approach to analyze tweets. In their paper, they try to assess the users' posts from a sample of six months according to various emotional attitudes: fear, hope, worry, happiness, anxiety, and nervousness by applying a Pearson correlation. It turns out that emotional tweets in general, as opposed to neutral tweets, are negatively related to the values of one-day-ahead stock indices such as DJIA, S&P 500 and NASDAQ, while positively related to VIX - a stock market expected volatility reported by Chicago Board Options Exchange. This result implies that at times of increased uncertainty, Twitter users tend to express themselves with emotional words, regardless of the context of uncertainty, whether it is positive or negative. The authors also try to use the total number of followers that might have seen the certain emotional tweets as well as the number of re-tweets of emotional tweets, as two other potential predictors of the values of indices. However, they happen to have less power compared to the number of emotional tweets.

Twitter is one of the most famous social networks nowadays, and it has an enormous impact on how people perceive and share information. Nevertheless, we would like to touch upon the influence of other websites such as Google and LiveJournal, as we believe it will provide us with a better understanding of how various social media and global networks can relate to economic and financial indicators.

Initially, we look into Google queries and any capability of predicting economic activity. Whether Google queries can aid in forecasting economic activity is the core of the study tackled by Choi and Varian (2009). The reason for choosing Google Trends data over governmental data, which is intensely used by numerous economists and investors, is due to the fact that government reports are merely obtainable with a lag since the data reports for a certain month are issued midway through the upcoming month and reviewed a number of months afterward.

However, Google Trends deliver weekly and even day-to-day reports on the volume of queries of different industries. In their study, Choi and Varian claim that Google Trends data can assist in predicting the present rather than the future. They assume that the possibility of having a correlation between query data and the present state of economic activity in a certain industry, might, in fact, support the prediction of the following data publication. An index of the volume of Google queries is conveyed by Google Trends. This index is sorted by geographic location and category. For a specified search word, Google Trends report a query index instead of raw level queries. Choi and Varian (2009) find that seasonal autoregressive models (AR) and fixed effect models that involve Google Trends variables, outperform models that lack these variables. 'Retail sales', 'automotive sales', 'home sales' and 'travel' are taken as study examples. In some cases, the gain reaches a significant amount such as the 18% improvement in the predictions for 'motor vehicles and parts' and the 12% improvement for 'new housing starts'.

In conjunction with the preceding, we look into the impact that emotions have on actual world states and individual's decisions. To what extent emotions influence real-world settings such as financial markets and how much people's emotions affect their choices, is the main concern of the study by Gilbert and Karahalios (2010). In this study, various emotional states such as anxiety, fear, and worry, are assessed in a dataset of over 20 million posts on the LiveJournal website. An index called Anxiety Index is constructed out of the metric of emotional states (anxiety, fear, and worry). According to Alan Greenspan, "Fear is an automatic response in all of us to threats to our deepest of all inbred propensities, our will to live. It is also the basis of many of our economic responses, the risk aversion that limits our willingness to invest and to trade, especially far from home, and that, in the extreme, induces us to disengage from markets, precipitating a severe falloff of economic activity." (Greenspan 2007, p. 17). Gilbert and Karahalios (2010) find that the Anxiety Index contains signals about future stock market prices that are not clearly visible from the market data. The authors identify a Granger Causal relation between an algorithmic estimation of the mood of millions of people, and the stock market. They find that the Anxiety Index has novel data about the S&P 500, approximately 70% of the 2008 trading year and that a single standard deviation increase in the Anxiety Index corresponds to 0.4% lower returns (its actual returns and not log returns). It is important to point out the big difference in return decrease when comparing the 0.07% to the 0.022% of Tetlock (2007) and Hirshleifer and Shumway (2003) respectively. The anticipation of this vast gap is due to the wideness of the scope of the Anxiety Index and the unexpected market swings of 2008 combined.

Three essential contributions that take place in the paper of Gilbert and Karahalios, are: tapping the emotions of a massive cluster of people without using a proxy such as mainstream media, validating that this technique delivers appropriate information about the stock market, and sticking exclusively to algorithmic techniques without accepting any human interferences. Using a Granger Causal framework, Gilbert and Karahalios find that rises in levels of anxiety, demonstrated by computationally identified linguistic structures, forecast downward force on the S&P 500 index. The results are also confirmed by the Monte Carlo simulation. The discoveries display how the mood of masses in an enormous online community can anticipate changes in an apparently unconnected system.

Therefore, we recognize that Twitter and other global networks have been gaining a lot of importance in the literature lately. There are multiple studies devoted to the subject of the public mood and its connection with various political, socio-economic and financial events. Even though a number of methods and research papers regarding Twitter already exist, we want to perform our own study in order to develop a method and a tool for evaluating the public's mood, using social media with Twitter as an example.

As a final article in the literature review, we genuinely explore machine learning, which is an existing application of Artificial Intelligence (AI). Machine learning is a field of computer science that uses a statistical approach to provide computer systems with the ability to learn. In other words, machine learning is about exposing models to new data and expecting them to self-sufficiently adapt. Thanks to new computing technologies, today's machine learning has evolved a lot and is adopted by various industries (financial services, healthcare, transportation, marketing, etc.), mostly by the industries that work with enormous amounts of human data. This vast adoption of machine learning is due to its ability to automatically implement complicated mathematical computations to massive amounts of data, repeatedly, during a remarkable time. A significant number of papers are written about machine learning.

In their paper, Naiknaware, Kushwaha, and Kawathekar carry on social media sentiment analysis using machine learning classifiers. The authors of the paper create a system, which collects, classifies and scores sentiments of tweets. Different methods, such as Maximum Entropy, Naïve Bayes and Support Vector Machines are applied during the experimentation. The performance of the classifiers is from seven datasets (Budget2017, Demonetization, GST2017, Digital India, Kashmir, Make in India, Startup India).

The best results are performed by SVM in ‘GST2017’, Naïve Bayes in ‘Budget2017’ and ‘Demonetization’ and Maximum Entropy in ‘Digital India’, ‘Kashmir’, ‘Make in India’ and ‘Startup shows’.

Accordingly, the relation between social networks and socio-economic indicators has been receiving lots of attention from a number of scientists. Conjointly, in our paper, we attempt to apply various techniques, including machine learning, to find a relation between public sentiment and stock returns. The detailed explanation of these methods will follow later in chapter 4.

3. Data: collection and general overview

For this study, we need two types of data: Twitter posts and stock returns.

3.1 Twitter data

Users' posts from Twitter need to be parsed in order to be further processed in our analysis. For this, we will use two main resources. The first one is a computer application called TAGS (Twitter Archiving Google Sheet), which is a devoted tool that helps in searching for the most recent tweets within the last six to nine days according to specific search inquiries. An example of this application process could be searching for 'cashtags' such as '$AAPL', which stands for Apple stock. The second resource is our own Twitter Crawler implemented in the Java environment.

The use of two sources at the same time is justified by the fact that there are significant limitations of the Twitter Search, Application Programming Interface (API). Every modern network has an API that allows anyone else to work with the website interface. It is mentioned on the Twitter platform for developers[1] that not all tweets are directly available whenever one uses the search engine. It has also been found that "the search API over-represents the more central users, and does not offer an accurate picture of peripheral activity" (González-Bailón, Sandra, et al., 2012). Central users are users with higher numbers of followers and with more popular Twitter profiles. Consequently, both sources complement each other; duplicated tweets are deleted at the end of the merging of the two datasets.

In the process of tweet cleaning, we clean 'spam' (for instance, deleting tweets featuring many 'cashtags', which is a clickable ticker symbol that gives a chance to quickly search for tweets and news about particular companies and stocks). We then identify 'buzz words' (such as FREE, ACTIVE TRADERS etc.) because tweets containing these words are considered irrelevant. In addition, we delete 'stop words' by using an extensive list of words (such as a, the, above, all, to, so, etc.), since such words do not have any grammatical or semantic value and do not add any value to the analysis. Within the process of cleaning, we also perform lemmatization (for example is, are, am, were, and was, are all transformed into the verb be, etc.), we transform CAPS LOCKED tweets into normal font, we normalize tweets (for instance, converting 'suuuuuuch a craaaaazy decrease' into 'such a crazy decrease') and we clean pictograms (better known as 'emojis').

[1] http://www.developer.twitter.com

After finishing the initial processing of tweets, we have 269 832 tweets corresponding to 489 companies from the S&P 500 index (out of 302 671 tweets initially parsed). However, since many companies are not widely represented in the sample, we are going to narrow down our sample to the 14 most presented companies. The criteria for selecting these 14 companies are having at least 3000 tweets per ticker during the period of the analysis as well as having a minimum of 8 tweets per ticker on a daily basis.

The reason for imposing these limits is to make the analysis more robust and to guarantee that we have enough textual data to work with. After the selection process, 181 776 tweets will be used in our analysis, with the sample covering a period from December 26th, 2017 until March 15th, 2018. The companies included in the sample are the following:

Table 1. The sample of 14 companies to be used in the analysis after the selection process

Company	Ticker	Industry	Number of tweets in the sample
Apple	AAPL	Technology Hardware, Storage & Peripherals	40 007
Amazon	AMZN	Internet & Direct marketing retail	31 101
Facebook	FB	Internet Software & Services	24 972
NVidia Corporation	NVDA	Semiconductors	15 130
General Electric	GE	Industrial Conglomerates	13 343
Alphabet	GOOGL	Internet Software & Services	9 336
Advanced Micro Devices	AMD	Semiconductors	8 667
Walmart	WMT	Hypermarkets & Super Centers	8 255
Alphabet	GOOG	Internet Software & Services	6 828
Bank of America Corp.	BAC	Diversified Banks	6 454
AT&T	T	Integrated Telecommunication Services	5 198
Netflix	NFLX	Internet Software & Services	5 086
McDonald's	MCD	Restaurants	4 096
Microsoft	MSFT	Systems Software	3 303

The number of tweets corresponding to each ticker, the weight of the company in S&P 500, the number of friends each user has and the number of re-tweets each tweet has, have all been added to the final dataset. This information may become useful when we start applying various machine learning techniques.

3.2 Financial data

We obtain the daily stock prices of S&P 500 companies from the Center for Research in Security Prices (CRSP) database provided by Wharton Business School.

Then, the daily stock log-returns are calculated according to the formula below:

$$R_t = ln\frac{p_t}{p_{t-1}}$$

where (p_t) is the closing price of the stock at a time 't', (p_{t-1}) is the price of the stock on the preceding day and (R_t) is the daily log-return.

In cases where there has been a tweet, but there has been no stock return - due to the fact that the stock market was closed on that day and there was no trading - we have merged that specific tweet with the closest stock return available within a three-day interval of the posting of the tweet. The final sample of stock returns as well covers a period from December 26th, 2017 until March 15th, 2018.

3.3 Preliminary summary of stock returns depending on the number of tweets

Wall Street's crash in early February 2018 is undoubtedly its wildest week since the 2008 financial crisis. The biggest single-day point decline in the Dow Jones entire trading history took place on the 5th of February, 2018. DJIA had plunged by 1 597 points, only to make a slight recovery by closing at 1 175 points, which is a decrease of 4.6%. This sharp plunge was sensed around the world reaching Europe and Asia. On the 6th of February, Hong Kong's Hang Seng took a 5.1% hit, while Japan's Nikkei fell by 4.7%. Moreover, major cities in Europe dropped by 2%, while Sweden, Germany, and Spain entered a correction.

Two days after, on the 7th of February, Donald Trump tweeted for the first time on this matter: "In the 'old days,' when good news was reported, the Stock Market would go up," and "Today, when good news is reported, the Stock Market goes down. Big mistake, and we have so much good (Great) news about the economy!" (Donald Trump).

The S&P 500 closed even slightly lower than the Dow on that day. After Donald Trump broke his silence, the very next day, DJIA reported the second-worst fall in history on the 8th of February of that same week. It plummeted by 1 033 points or 4.2%.

On the 9th of February, the Dow Jones index closed up 330 points. However, it was not obvious which direction stocks were moving. At one point in time, DJIA was down by 500 points, at another, it was up by 500 points. In just two weeks, the Dow had plunged more than 3 200 points or 12%.

According to numerous researchers, Twitter is capable of predicting stock prices more accurately than any other investment tactic. A team of researchers at the University of California has invented a computer model, which allows them to forecast the stock market by scanning the social network.

The team claims that one can discreetly predict what could happen in the stock market tomorrow, by simply using Twitter. Their model has up to 11% more accuracy than other computer models. It even outperformed the Dow Jones Industrial Average (DJIA).

At this level, we are testing a hypothesis that the sharp changes in the number of tweets might be capable of predicting fluctuations in the stock market. To do so, we will conduct a case study on the S&P 500 firms by looking into the number of tweets and at the (t+1) log-returns over the period between January 2nd and March 14th, 2018.

Our main point of interest is to see whether there has been any sharp increase in the number of tweets on the week preceding the 5th of February and if there was one, to compare the volatility of stock returns one month before this event and one month after it.

After applying this study method to the top ten companies in the list, based on both the share in the S&P 500 total market capitalization and the total number of tweets available in our sample, we will see whether there are certain patterns that we can observe in the behavior of stock returns before and after the event.

If there is some kind of relation that we can recognize, but there are also companies that fall out of the pattern, then we need to dive deeper into our analysis of Twitter. In order to do that, we would need to consider not only the number of tweets corresponding to certain stock tickers but also the contents of tweets. This approach will help us to understand how certain kinds of information can be correlated with the future stock returns. The method is called Sentiment Analysis; it is what we are going to describe in the methodology part and what we will further implement at the next stage of our paper.

After looking into the top companies in our sample, and building graphs depicting how the stock returns were behaving in January and February (referring to Appendix 1), we can recognize a strong pattern. In the period between 29th of January and 1st of February, there has been a sharp increase in the number of tweets corresponding to the following six business ‘giants’: Facebook, Google, Amazon, JPMorgan, McDonald’s and Microsoft.

After that, the S&P 500 encountered a very volatile period we have previously described. In the table below, we outline the sharp change in return volatility that occurred to these stocks.

Table 2. Quantified increase in stock volatility between January and February 2018 for the selected companies with the largest number of tweets available

Company	Stock return volatility (st. dev.) - January 2018	Stock return volatility (st. dev.) - February 2018	Change
Facebook	0,0125	0,0215	**+72%**
Google	0,0076	0,0237	**+212%**
Amazon	0,0155	0,0249	**+61%**
JPMorgan	0,0060	0,0215	**+259%**
McDonald's	0,0085	0,0173	**+104%**
Microsoft	0,0090	0,0225	**+150%**

This is bringing us to the conclusion that a significant increase in the number of tweets, can serve as an indicator of public panic, which in turn might partially account for the changes in stock return volatility. In order to dive deeper into this issue, we now need to examine the contents of tweets to understand whether it matters and whether the positive/negative sentiments encapsulated in the tweets can cause certain shifts in the stock markets. One way to check this relation is to apply sentiment analysis to the tweets by classifying their semantics into positive, neutral and negative. Another method to go around this issue is to implement more sophisticated machine learning techniques.

4. Methodology

4.1 Lexicon-based sentiment analysis of Twitter

The first step in the analysis of the relation between Twitter and financial markets is a Lexicon-based approach. We apply two different Lexicon-based methodologies to calculate the sentiment scores of each tweet.

The first one is based on the method introduced by Finn Årup Nielsen[2], who created a large lexicon corpus that includes a variety of words and phrases ranked between -5 (extremely negative) and +5 (extremely positive)[3]. The final unweighed score is calculated by summing up the scores of words or phrases in every tweet. After that, for every stock ticker at each point of time (daily) we weigh the individual tweet scores by the number of followers that a particular user has, and we obtain the panel dataset of weighed tweet sentiment scores.

The second method is based on the first one, but with an additional twist. First, after obtaining the raw scores of every tweet, we normalize the values by assigning to every tweet either a positive (+1), neutral (0) or negative (-1) semantic value based on the kind of information contained in the post. As a result, we obtain a panel dataset of classified tweets for every day in the sample, for each company. Then, we calculate the following sentiment index:

$$SE_{it}\ (Sentiment\ Index) = \frac{tw_{it,+} - tw_{it,-}}{tw_{it,+} + tw_{it,-} + tw_{it,0}}$$

where ($tw_{it,+}$) is the number of positive tweets, ($tw_{it,-}$) is the number of negative tweets and ($tw_{it,0}$) is the number of neutral tweets for each company 'i' at a time 't'.

After the processing of raw tweets is over, and the final sentiment score of both methodologies is calculated, we apply a simple time-series regression for each ticker using two various separate series of calculated sentiment scores:

$$Stock\ Return_t = \beta_0 + \beta_1 Sentiment\ Score_{t-1} + \varepsilon_t$$

The reason behind using time-series regressions for each company individually is that different companies might exhibit various degrees of sensitivity to Twitter activity; some companies differ in terms of showing positive/negative relations.

[2] AFINN sentiment analysis in Python: https://github.com/fnielsen/afinn

[3] The process of classifying the data in the tweets is applied in the R programming environment.

In addition to this, we pool all the companies together and implement a simple panel data regression for the following model:

$$Stock\ Return_{it} = \beta_0 + \beta_1 Sentiment\ Score_{i(t-1)} + \varepsilon_{it}$$

There are two main advantages of panel data over time-series regression. The first advantage is the ability of panel data to control for unobserved heterogeneity and the second is the enhanced accuracy of conclusions due to a higher number of degrees of freedom, thanks to the two data dimensions.

After applying this methodology to the stock returns, we decide to implement the second method, in which we:

1. Clean the company stock returns from the market component by implementing the CAPM model (for the time-series only, as the panel data model cannot be estimated for the CAPM):

$$Excess\ Stock\ Return_t = \beta_0 + \beta_1 Excess\ Market\ Return_t + \varepsilon_t$$

2. Take the residuals ε_t from the respective regressions, use them as a proxy for the unexplained-by-the-market residuals of stock returns and insert them in the following time-series regression:

$$\varepsilon_t = \beta_0 + \beta_1 Sentiment\ Score_{t-1} + u_t$$

4.2 Machine learning-based sentiment analysis of Twitter

The process of classifying tweets into labels is not that simple, given that sometimes we have to determine which sentences can serve as a buy/sell signal. Sometimes, the lexicon-based approach mentioned above misses the complicated connections between words in sentences. That is why we resort to artificial intelligence and machine learning techniques in particular to resolve this issue.

Artificial intelligence, which is a broader term including, but not limited to machine learning only, is the ability of computers to mimic human activity as in perceiving and learning information, reasoning, and problem-solving. Machine learning is comprised of various algorithms, which process large sets of data to find dependencies and relations to provide the user with the ultimate predictions and recommendations based on the uncovered patterns. Machine learning is widely used in many spheres of our lives, from medicine and finance to art and linguistics. It has an enormous power in both scientific and practical terms.

Machine learning can be either unsupervised or supervised. The main difference between the two is that under the former one, the model is not given the desired output, i.e. there is no ultimate goal set for the algorithm. Therefore, the unsupervised machine learning is predominantly used in clustering data.

In our analysis, we apply supervised machine learning as we provide the model with the output variables (the stock return flag). Supervised machine learning is widely used in text classification tasks.

We should also mention the concept of Natural Language Processing (NLP) - a field of computer science, which explores algorithms for processing natural language corpora, establishing interactions between computers and humans. NLP studies the diversity and complexity of elementary units (for instance, words) and the corresponding structures (for instance, sentences or articles) formed by uncountable combinations of these elements. Aiming to understand which information tweets tolerate and how it can be interpreted in terms of affecting stock returns, we will inevitably touch upon various applications of NLP in our essay.

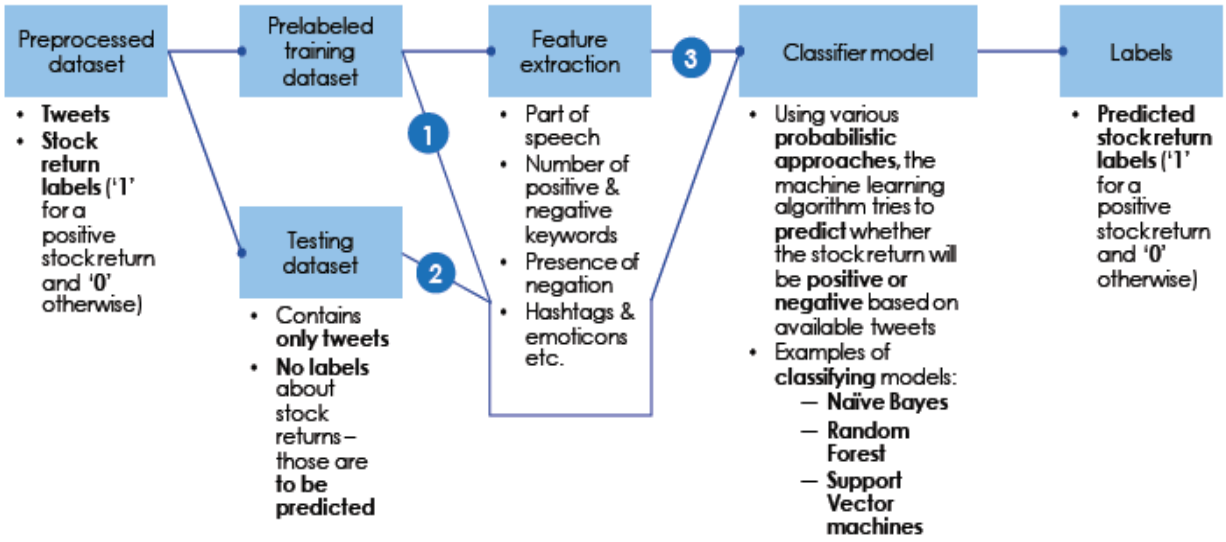
In the scheme below describing the machine learning process, the procedure is initiated by dividing the initial dataset, which is comprised of preprocessed Twitter posts and daily stock return labels for each company, where the stock return is labelled with '1' if there has been an increase in stock price observed, and with '0' otherwise.

The first part of the initial dataset is called training dataset and it includes 75% of the initial dataset (the period from December 26th, 2017 until February 23rd, 2018). The training dataset is used for feature extraction, where machine learning algorithms learn various features from textual information contained in tweets; this includes parts of speech, number of positive/negative keywords, the presence of negations and positive/negative hashtags/emojis. It then finds the relation between various words, phrases, features, and labels, which all these have resulted in.

After the process of learning is finished, we take the machine learning algorithm - the features found in the training dataset and apply them to the testing dataset, which constitutes of 25% of the initial dataset (the period from February 24th, 2018 until March 15th, 2018). This includes only preprocessed tweets but does not include any labels regarding an increase/decrease in stock prices. All of this is done through a classifier model, which uses various statistical approaches to predict whether the stock return is going to be positive (label of 1) or negative (label of 0) based on the tweets available. In this paper, we will use four different classifier models, which we will touch upon in more detail further on.

In the final stage, once we obtain the predicted labels for stock returns, we compare them with the actual stock return labels and calculate the accuracy rate of the model by calculating the number of correctly predicted labels. This is the ultimate measure, which will indicate whether these machine learning models can have any predictive power.

Figure 1. Machine learning approach to sentiment analysis of Twitter



4.3 Classifier models used in the machine learning approach

As it has been mentioned earlier, the machine learning algorithms rely on certain statistical models, which use the learned features and help predict the stock return labels. We will cover four different classifier models:

- Naïve Bayes
- Random Forest
- Support Vector Machines (Gaussian)
- Support Vector Machines (Linear)

4.3.1 Naïve Bayes classifier

As the name of this model signifies, Naïve Bayes classifier relies on the Bayes theorem, which points to how the probability of the event can be described based on the information about other conditions related to the event. Translating this to the language of our topic, the probability of assigning a certain label to the stock return can be described based on the information regarding the label that has been assigned earlier given the same or similar features contained in the tweets (Lewis and Gale, 1994).

First, we calculate the probabilities of all the labels given the textual information, and then, we select the label with the highest probability. This definition can also be written mathematically as:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c|t)$$

$$\boldsymbol{P(c|t) \propto P(c) \prod_{i=1}^{n} P(f_i|c)}$$

The Naïve Bayes method considers all features differently, which is often perceived as a drawback. However, it has also turned out to be one of the most efficient classifiers.

4.3.2 Random Forest classifier

The Random Forest is an ensemble technique that combines multiple decision trees to predict the most likely label to be assigned, based on the given data (Breiman, 2011). The decision tree itself is a simpler algorithm, which drills down each sentence asking various questions based on words and phrases contained in the tweets. Each node is divided into two branches each time (based on whether the answer on the last question was 'true' or 'false'), and the process continues until the decision tree can give a definitive answer, giving the label a '0' or a '1'. The criteria for selecting questions are based on all features learned and information gained, which is calculated through complicated formulas. An example of a decision tree for our context would be an algorithm consecutively asking questions such as: 'Is there a negation in the sentence?', 'Does the sentence contain the word 'sharp decline'?', 'Are there any emoticons in the tweet?' and so on.

After each decision tree reaches its end, they are pooled into one model, the Random Forest. The reason for doing so is because each predicted value we get in any decision tree is bound to having a high variance that might not be accurate. Random pooling of multiple decision trees helps in overcoming this problem. Running a random forest model, instead of a decision tree model, is comparable to base an investment decision on asking 100 finance professionals instead of consulting only one.

4.3.3 Support Vector Machines classifier (Linear)

Support Vector Machines is based on the principle of finding a hyperplane that would best divide all the data points into two classes. The goal of the method is to maximize the distance between that hyperplane and the data points closest to it (Vapnik, 1995).

These data points are called Support Vectors; in case one deletes them from the sample, the hyperplane will change its positioning. This method is easily trained and is applicable when it is possible to linearly divide the data space.

4.3.4 Support Vector Machines classifier (Gaussian)

However, sometimes it is hard to divide the dataset into two separate parts since all the observations look mixed up with each other. In this case, the procedure of kernelling is applied. For example, we add another dimension to the 2D view and examine the data points in the 3D space to see if we can find a hyperplane to fit in. This procedure continues (by moving into higher and higher dimensionality) until we find a suitable hyperplane. This is called a non-linear kernel in the Support Vector Machines approach (Sung, 1995). We will be using the Gaussian kernel.

5. Results

5.1 Lexicon-based approach: time-series model

Before applying any time-series regressions to the data we collected, it is important to make sure that we are not dealing with unit-roots in the dependent variables. In other words, we need to check our dataset for the presence of stationarity, i.e. all the dependent variables have a constant mean, variance, and auto-covariance. In case our data contains random walks or processes with exploding variances, we will encounter spurious regressions and our results would be completely irrelevant.

To perform the test of stationarity, we implement the Augmented Dickey-Fuller (ADF) test, in which the null hypothesis is the presence of unit-roots. We apply this test to the two time-series variables: stock returns and residuals of stock returns obtained after running the CAPM. Even though stock returns should be stationary by default, since they are the first differences of stock prices, we still perform the test on them to make sure we are not vulnerable to spurious regression inferences. The full STATA outputs can be found in table 4 of Appendix 2. In table 6 of Appendix 2, we present the summary of the ADF test results that we have obtained.

By referring to table 4 in Appendix 2, we can see that both stock returns and residuals of stock returns are stationary for 13 out of 14 companies in our sample, hence, they can be used in the analysis.

In order to validate the results of the Dickey-Fuller test, we apply a Kwiatkowski Phillips Schmidt Shin (KPSS) test. The null hypothesis in the KPSS is the opposite of the ADF test, thus, all the variables are assumed to be trend-stationary according to H_0. The test is as well implemented in the STATA environment and is applied to the two time-series variables we are working with at this point, which are: stock returns and residuals of stock returns obtained after running the CAPM. Hence, it is safe to use both variables, after excluding the ticker WMT (Walmart) for which this variable turned out as non-stationary. The full STATA outputs can yet again be found in table 5 of Appendix 2. The summary of the KPSS test results that we have acquired is presented in table 7 of Appendix 2.

Table 7 of Appendix 2 indicates that KPSS proves the conclusions we have reached about stationarity for the majority of tickers. Therefore, it is safe to continue to use both variables, after excluding three tickers: GE (General Electric), MCD (McDonald's), and T (AT&T). The final sample for time-series analysis includes the following companies, with a total of 150 884 tweets available:

Table 3. Final sample of companies included in the time-series regression analysis

Company	Ticker	Industry	Number of tweets in the sample
Apple	AAPL	Technology Hardware, Storage & Peripherals	40 007
Amazon	AMZN	Internet & Direct marketing retail	31 101
Facebook	FB	Internet Software & Services	24 972
NVidia Corporation	NVDA	Semiconductors	15 130
Alphabet	GOOGL	Internet Software & Services	9 336
Advanced Micro Devices	AMD	Semiconductors	8 667
Alphabet	GOOG	Internet Software & Services	6 828
Bank of America Corp.	BAC	Diversified Banks	6 454
Netflix	NFLX	Internet Software & Services	5 086
Microsoft	MSFT	Systems Software	3 303

Now we are ready to perform the time-series regression analysis by running a simple linear time-series regression in STATA.

The coefficients, t-stats and adjusted r-squared obtained for each ticker of the original stock returns model can be found in tables 8.1 and 8.2 of Appendix 2, while the coefficients, t-stats and adjusted r-squared acquired for each ticker of the residual stock returns model can be found in tables 9.1 and 9.2 of Appendix 2.

All four tables, 8.1, 8.2, 9.1, and 9.2 of Appendix 2 show that no statistically significant relation between stock returns and sentiment scores based on tweets could be found (neither for the models with original stock returns nor for the models with residuals of stock returns after the CAPM application).

5.2 Lexicon-based approach: panel data model

Before implementing the panel data model, we need to conduct several tests. The first one is the Hausman test, which will help us to choose between the Random Effects model (RE) and the Fixed Effects model (FE). We execute this test for the regression with the original stock returns. The independent variable will be either the normalized or the weighed sentiment score. In tables 10.1 and 10.2 of Appendix 2, we can see the outputs of both cases (normalized and weighed sentiment score). The results of both test outputs signify that we should use the random effect models since the p-values are higher than 0.05.

The next test is the Heteroscedasticity test. This test is conducted using maximum likelihood and is implemented in STATA. The outputs of both cases (normalized and weighed sentiment score) can be found in tables 11.1 and 11.2 of Appendix 2. The results of the Heteroscedasticity test point out that the null hypothesis of homoscedasticity is not rejected for either models. Consequently, we do not need to include robust standard errors in the model.

The final test is the Autocorrelation test. The STATA package, by David Drucker, is used for implementation. The outputs of both cases (normalized and weighed sentiment score) can be found in tables 12.1 and 12.2 of Appendix 2. From the results of the Autocorrelation test, we can observe that both models contain serial autocorrelation. Accordingly, we will account for it by using robust clustered standard errors.

At this point, after we have conducted all the preliminary tests, we can run the relevant panel data models. The results can be found in tables 13.1 and 13.2 of Appendix 2. As we can see from the regression outputs of the panel data models, the coefficients for normalized and weighed sentiment scores are significant in both models. In addition, the models themselves are also showing overall significance, since the statistics for Wald chi2 is statistically significant. The coefficients we obtained, signal that if the normalized or weighed sentiment score is expected to increase by 1 point today (by 0.01 since our sentiment scores are measured on the scale from 0 to 1), then the next day the stock return for that company is expected to increase by 0,0082% in the case of normalized sentiment score and by 0,0017% in the case of weighed sentiment score. If we take Apple ($AAPL) for instance and assume that the weighed sentiment score increases by 1 point, its total market capitalization could see an increase of $15.6 million the next day.

In order to conduct the robustness check, we now get back to the full dataset including all the companies we have data for (489 firms of the S&P 500 index). We run both models (normalized and weighed sentiment scores) with the inclusion of clustered standard errors and using the RE (random effects) model. The full regression outputs can be found in tables 14.1 and 14.2 of Appendix 2. The results indicate that using the full dataset instead of a limited, but with a more balanced one (with the selected 14 companies only) lowers the significance of the results. Thus, we can conclude that thoroughly choosing the sample of firms has helped us to attain better regression outputs.

5.3 Machine learning approach

Now we are moving to a more complicated technique, which involves the machine learning algorithms. We have implemented the four earlier described models using the three following methods:

- Unigram
- Bigram
- Trigram

The difference between these three models is in the number of words each model considers as a source for learning features and connecting them to stock returns. The algorithm uses separate words in the case of unigram, pairs of words in the case of bigram and three consecutive words in that of the trigram.

Refer to table 15 of Appendix 3 for the output of unigram, table 16 of Appendix 3 for the output of bigram and table 17 of Appendix 3 for the output of trigram. All three outputs of unigram, bigram, and trigram indicate that the random forest and the SVM-Gaussian have the highest prediction rates on average, which makes them the best models to predict the stock returns based on tweets using any of the three methods (unigram, bigram, or trigram).

The percentages presented in the unigram, bigram, and trigram tables act as a measure of the accuracy of each method. For example, this means that there is a 68% probability that the unigram Naïve Bayes method will predict the up/down movement of the Amazon stock return correctly, there is a 64% probability that the bigram Random Forest method will predict the up/down movement of the Netflix stock return correctly, and there is a 60% probability that the trigram SVM-Gaussian method will predict the up/down movement of the NVidea stock return correctly.

6. Conclusion

As we have already mentioned in the literature review and as a series of previous papers have publicized, there has been a signal worth investigating that connects public sentiments in social networks such as Twitter to market behavior.

This paper's prime aim was to extend additional empirical research in contemplation of further investigating the impact of Twitter sentiments in tweets on the stock market movements, and methodically testing whether the overall public sentiment of tweets is effectual in forecasting the stock returns of S&P 500 companies.

The data we needed for our thesis consisted of Twitter data and financial data; our Twitter posts were collected using Twitter Archiving Google Sheet (TAGS) in addition to our own Twitter Crawler implemented in the Java environment, and our daily stock prices of S&P 500 companies were acquired from the Center for Research in Security Prices (CRSP). After we cleaned the Twitter data and selected the companies with the highest amount of tweets, we had 14 companies containing 181 776 tweets between the period of December 26th, 2017 and March 15th, 2018 to work on in our analysis.

In our dissertation, we used the lexicon-based approach and various machine learning classifiers (Naïve Bayes, Random Forest, SVM-Gaussian, and SVM-Linear) for the test classification of tweets. After that, we tested the relation between stock returns and tweet sentiment scores. This was accomplished using regression analysis (implemented in the STATA environment) and machine learning.

Our thesis results demonstrated how the lexicon-based time-series regression, and after conducting two stationarity tests (ADF and KPSS tests) applied in STATA, failed to give any significant results on the relation between tweet sentiment scores and stock returns. However, we managed through the lexicon-based panel data regression, and subsequent to executing various tests (Hausman test for Random Effects/Fixed Effects, Likelihood ratio test for Heteroscedasticity and the Wooldridge test for Autocorrelation) applied in STATA, to present significant evidence of dependence between tweet sentiments and stock price returns of companies included in the S&P 500 index.

Moreover, in testing the relation using the machine learning approach, we implemented four different machine learning techniques: Naïve Bayes, Random Forest, SVM-Gaussian and SVM-Linear, where all four models were executed via three different methods: unigram, bigram, and trigram.

The percentages in our thesis results of all three tables (unigram, bigram, and trigram), indicated that the random forest and the SVM-Gaussian were the best models in predicting the stock returns based on tweets.

This work, based on advanced empirical techniques, presented compelling proof that a significant dependence relation between tweet sentiments and stock returns solidly stands, and a prediction power of stock returns based on tweets strongly exists.

With Twitter being capable of collecting vital information that resembles public moods and opinions, this dissertation can be utilized in building an effectual concrete system, which aids in constructing future investment strategies around forecasted stock returns for individual investors and high-frequency trading firms.

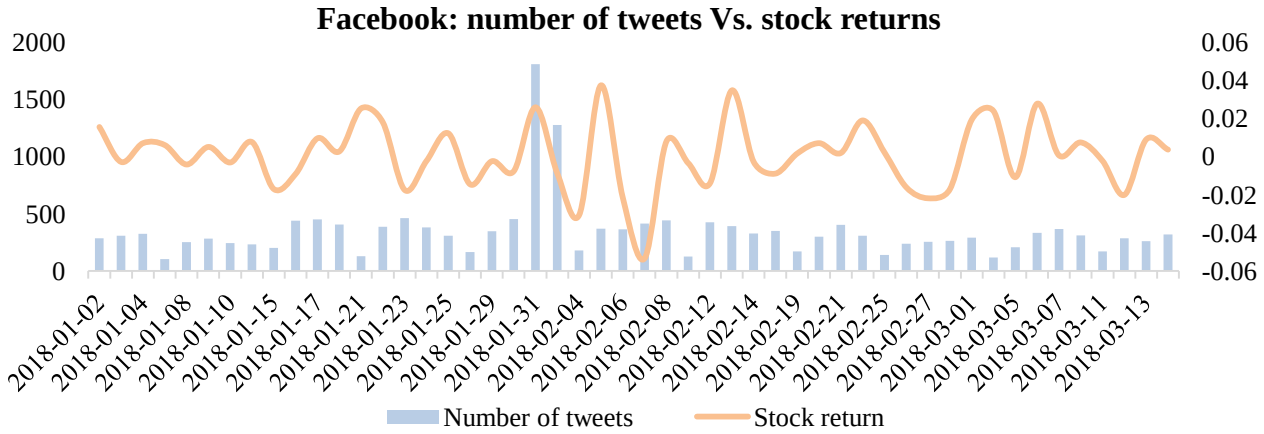
References

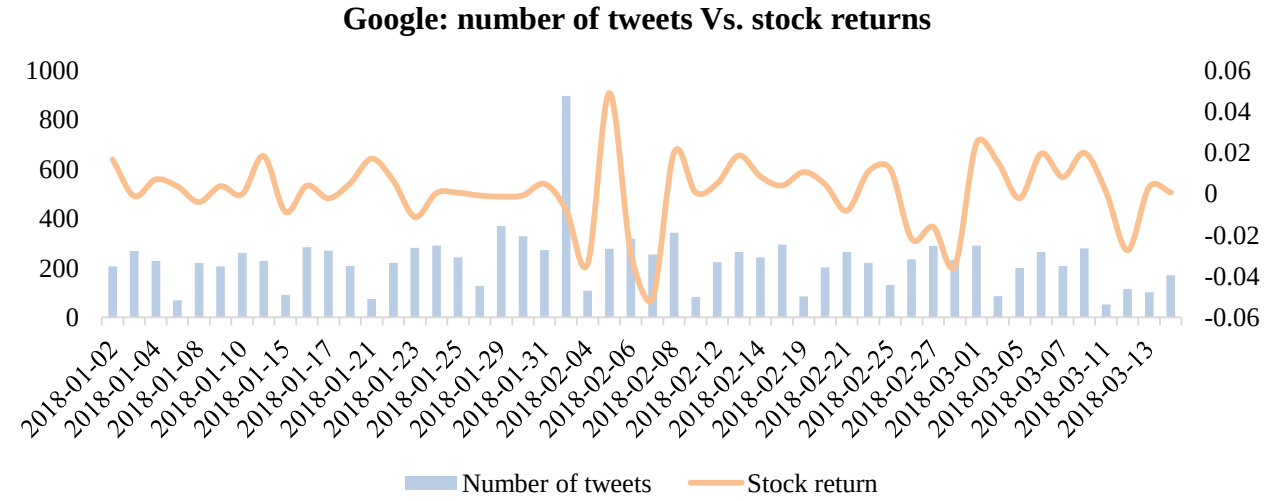
1. Asur S., Huberman B. A. Predicting the future with social media //Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. – IEEE, 2010. – Т. 1. – С. 492-499.
2. Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market //Journal of computational science. – 2010. – Т. 2. – №. 1. – С. 1-8.
3. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
4. Choi H., Varian H. Predicting the present with google trends. Google Inc. - 2009
5. Fama E. F. The behavior of stock-market prices //The Journal of Business. – 1965. – Т. 38. – №. 1. – С. 34-105.
6. Gilbert E., Karahalios K. Widespread Worry, and the Stock Market //ICWSM. – 2010. – С. 59-65.
7. González-Bailón S., Wang N., Rivero A., Borge-Holthoefer J., Moreno Y. Assessing the Bias in Samples of Large Online Networks. - 2012.
8. Hirshleifer D., Shumway T. Good Day Sunshine: Stock Returns and the Weather // The Journal of Finance Vol. 58, No. 3. - 2003. - С. 1009-1032
9. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In SIGIR-94, 1994.
10. Lo, Andrew W. and A. Craig MacKinlay, A Non-Random Walk Down Wall Street, (Princeton: Princeton University Press, 1999).
11. Lo, Andrew W., Harry Mamaysky and Jiang Wang (2000), "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation," Journal of Finance, 55, 1705-1765.
12. Liu L., Shu H., Wei J. The Impacts of Political Uncertainty on Asset Prices: Evidence from a Natural Experiment - 2014.
13. Malkiel B., Mullainathan S., Stangle B. Market Efficiency versus Behavioral Finance. Journal of Applied Corporate Finance, 2005, vol. 17, issue 3, 124-136
14. Nofsinger J. R. Social mood and financial economics //The Journal of Behavioral Finance. – 2005. – Т. 6. – №. 3. – С. 144-160.
15. Pastor L., Veronesi P., 2013a. Political uncertainty and risk premia// Journal of Financial Economics 110, 520-545.
16. Ranco G. et al. The effects of Twitter sentiment on stock price returns //PloS one. – 2015. – Т. 10. – №. 9. – С. e0138441.

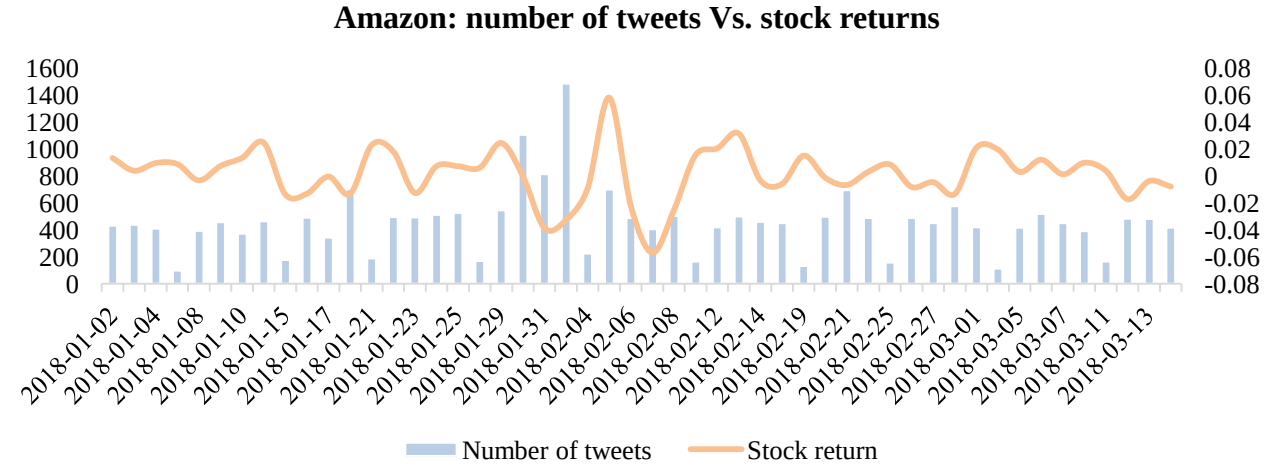
17. Shefrin, H. Behavioral finance: biases, mean-variance returns, and risk premiums. CFA Institute Conference Proceedings, Vol. 24, issue. 2 - 2007. - p. 4–12
18. Sung, K. 1995. Learning and Example Selection for Object and Pattern Detection. Ph.D. Thesis, Massachusetts Institute of Technology.
19. Tetlock P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. // The Journal of Finance. Vol. LXII, No 3. - 2007 - C. 1139-1168
20. Tumasjan A., Sprenger T. O., Sandner P. G., Welpe I. M. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape // Social science computer review 29 (4). - 2011. - 402-418
21. V. Vapnik, "The nature of statistical learning theory," Springer-Verlag: New York, 1995.
22. Zhang X., Fuehres H., Gloor P. A. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear" //Procedia-Social and Behavioral Sciences.– 2011. – T. 26. – C. 55-62.

Appendices

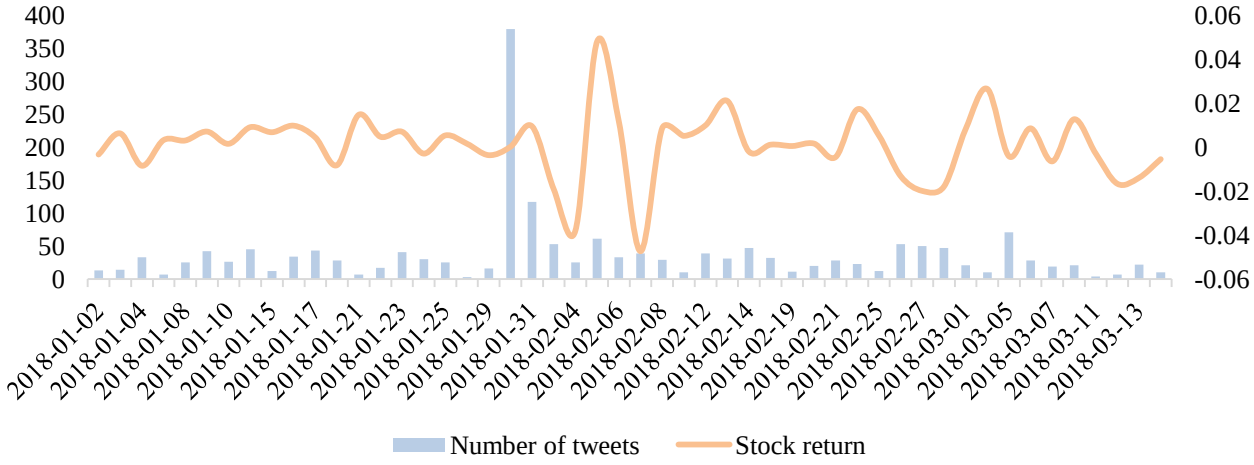
Appendix 1: Relation between the stock return volatility increase in February and the increased number of tweets at the end of January - beginning of February



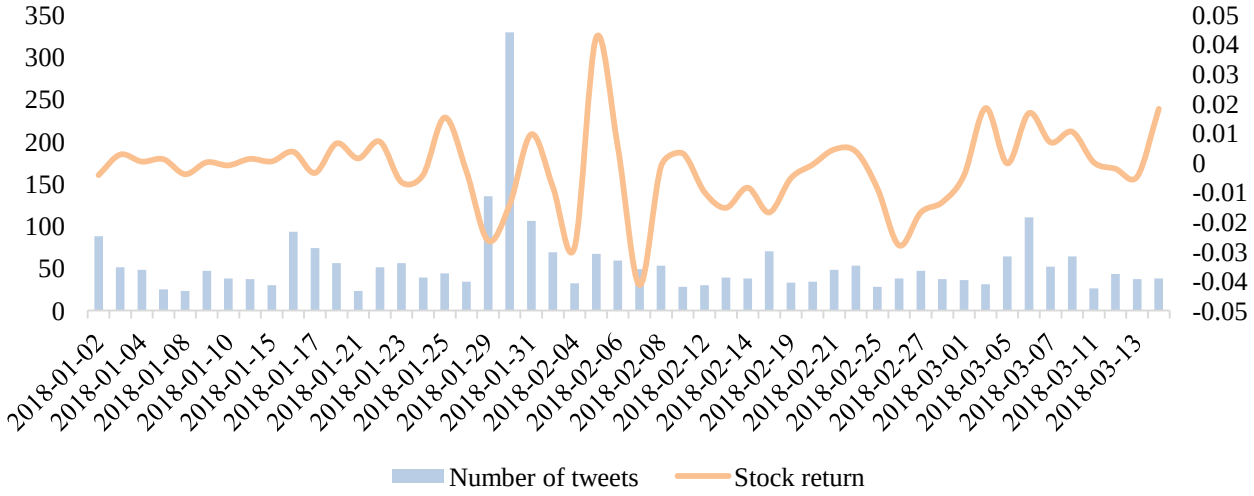




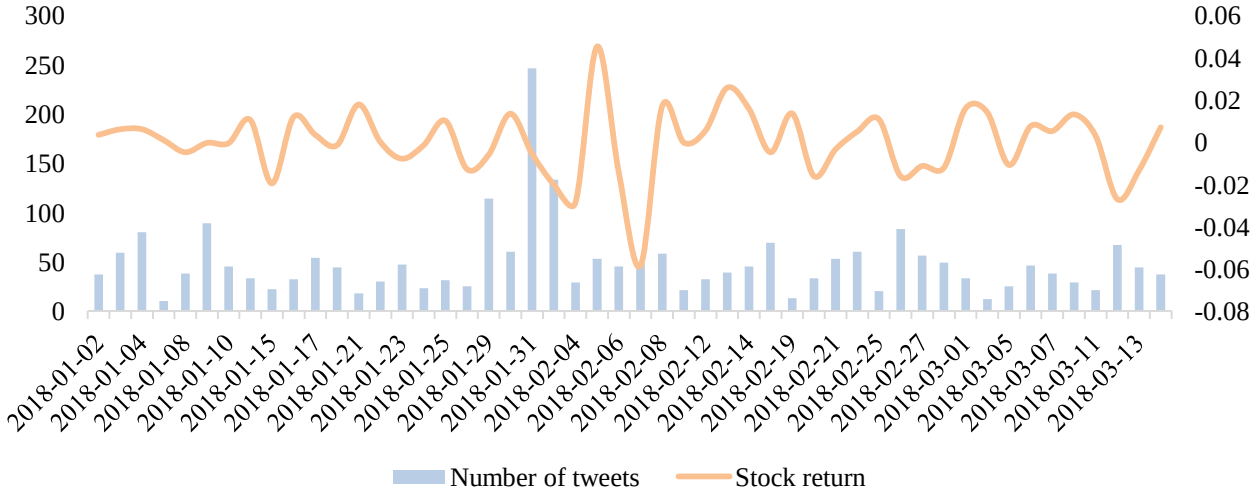
JPMorgan: number of tweets Vs. stock returns



McDonald's: number of tweets Vs. stock returns



Microsoft: number of tweets Vs. stock returns



Appendix 2: Regression analysis

Table 4. Augmented Dickey-Fuller (ADF) test for stationarity of time-series variables, H_0: the time-series variables contain unit-roots (the critical value is -2,591)

Ticker	Stock returns	Residuals of stock returns (after CAPM application)
	T-statistic	T-statistic
AAPL	-2,673*	-2,829*
AMD	-2,718*	-2,882**
AMZN	-2,627*	-2,666*
BAC	-3,169**	-3,286*
FB	-3,367**	-3,317**
GE	-2,751*	-2,852*
GOOG	-3,117**	-3,184**
GOOGL	-3,179**	-3,222**
MCD	-2,798*	-2,713*
MSFT	-2,709*	-2,789*
NFLX	-3,285**	-3,387**
NVDA	-2,8*	-2,866**
T	-3,262**	-3,306**
WMT	-2,257^	1,975^

* p<0.1, **p<0.05, ***p<0.01 : significance of t-statistic
^We fail to reject the H_0, thus, we consider these time-series variables non-stationary.

Table 5. KPSS test for stationarity of time-series variables, H_0: the time-series variables are stationary (the critical value is 0,216)

Ticker	Lag order	Stock returns	Residuals of stock returns (after CAPM application)
AAPL	0	0,162	0,156
AAPL	1	0,125	0,125
AAPL	2	0,115	0,118
AAPL	3	0,12	0,117
AMD	0	0,112	0,0989
AMD	1	0,0915	0,0842
AMD	2	0,0872	0,0826
AMD	3	0,0893	0,0859
AMZN	0	0,0767	0,0713
AMZN	1	0,0571	0,0545
AMZN	2	0,0546	0,0533
AMZN	3	0,0557	0,055
BAC	0	0,044	0,0452
BAC	1	0,0376	0,0401
BAC	2	0,0403	0,0438
BAC	3	0,0435	0,0474
FB	0	0,183	0,209
FB	1	0,147	0,166
FB	2	0,134	0,15
FB	3	0,125	0,139
GE	0	0,226*	0,225*
GE	1	0,162	0,166
GE	2	0,144	0,149
GE	3	0,137	0,141
GOOG	0	0,11	0,109
GOOG	1	0,0948	0,0966
GOOG	2	0,0923	0,0963
GOOG	3	0,0879	0,0925
GOOGL	0	0,124	0,128
GOOGL	1	0,105	0,111
GOOGL	2	0,101	0,107
GOOGL	3	0,0948	0,102
MCD	0	0,223*	0,244*
MCD	1	0,177	0,195
MCD	2	0,175	0,192
MCD	3	0,17	0,184
MSFT	0	0,084	0,0875
MSFT	1	0,0705	0,0737
MSFT	2	0,0718	0,0755
MSFT	3	0,0704	0,0744
NFLX	0	0,0746	0,069

NFLX	1	0,058	0,055
NFLX	2	0,054	0,0517
NFLX	3	0,0521	0,05
NVDA	0	0,0997	0,0929
NVDA	1	0,0825	0,08
NVDA	2	0,0834	0,0828
NVDA	3	0,0847	0,0848
T	0	0,21*	0,221*
T	1	0,152	0,18
T	2	0,0133	0,169
T	3	0,124	0,159

*The null is rejected. The series is non-stationary.

Table 6. Augmented Dickey-Fuller (ADF) test results summary, H_0: the time-series variables contain unit-roots

Ticker	Stock returns	Residuals of stock returns (after CAPM)
AAPL	H_0 rejected	H_0 rejected
AMD	H_0 rejected	H_0 rejected
AMZN	H_0 rejected	H_0 rejected
BAC	H_0 rejected	H_0 rejected
FB	H_0 rejected	H_0 rejected
GE	H_0 rejected	H_0 rejected
GOOG	H_0 rejected	H_0 rejected
GOOGL	H_0 rejected	H_0 rejected
MCD	H_0 rejected	H_0 rejected
MSFT	H_0 rejected	H_0 rejected
NFLX	H_0 rejected	H_0 rejected
NVDA	H_0 rejected	H_0 rejected
T	H_0 rejected	H_0 rejected
WMT	H_0 not rejected	H_0 not rejected

Table 7. Kwiatkowski Phillips Schmidt Shin (KPSS) test results summary, H_0: the time-series variables are stationary

Ticker	Stock returns	Residuals of stock returns (after CAPM)
AAPL	H_0 not rejected	H_0 not rejected
AMD	H_0 not rejected	H_0 not rejected
AMZN	H_0 not rejected	H_0 not rejected
BAC	H_0 not rejected	H_0 not rejected
FB	H_0 not rejected	H_0 not rejected
GE	H_0 rejected	H_0 rejected
GOOG	H_0 not rejected	H_0 not rejected
GOOGL	H_0 not rejected	H_0 not rejected
MCD	H_0 rejected	H_0 rejected
MSFT	H_0 not rejected	H_0 not rejected
NFLX	H_0 not rejected	H_0 not rejected
NVDA	H_0 not rejected	H_0 not rejected
T	H_0 not rejected	H_0 rejected

Table 8.1 Regression output for individual time-series regressions: dependent variable - stock returns, independent variable - normalized sentiment score

Ticker	Normalized Sent. Sc. Coeff.	Adjusted R-sq	# observations
AAPL	0,0334 (1,85)	0,0296	80
AMD	0,016 (0,83)	0,0088	80
AMZN	0,0178 (0,69)	0,0061	80
BAC	0,0105 (0,73)	0,0067	80
FB	-0,0154 (-0,58)	0,0042	80
GOOG	0,0135 (0,75)	0,0072	80
GOOGL	0,0092 (0,48)	0,0029	80
MSFT	0,0081 (0,71)	0,0064	80
NFLX	-0,0041 (-0,2)	0,0005	80
NVDA	-0,0159 (-0,54)	0,0037	80

Table 8.2 Regression output for individual time-series regressions: dependent variable - stock returns, independent variable - weighed sentiment score

Ticker	Weighed Sent. Sc. Coeff.	Adjusted R-sq	# observations
AAPL	0,0012 (0,42)	0,0023	80
AMD	0,0021 (0,43)	0,0024	80
AMZN	0,0037 (0,82)	0,0086	80
BAC	0,0029 (0,97)	0,0118	80
FB	0,005 (0,89)	0,0101	80
GOOG	0,0029 (0,84)	0,009	80
GOOGL	0,0003 (0,07)	0,0001	80
MSFT	-0,0015 (-0,56)	0,004	80
NFLX	0,0051 (0,99)	0,0125	80
NVDA	-0,0043 (-0,79)	0,0079	80

Table 9.1 Regression output for individual time-series regressions: dependent variable - residual stock returns, independent variable - normalized sentiment score

Ticker	Normalized Sent. Sc. Coeff.	Adjusted R-sq	# observations
AAPL	0,0346 (1,93)	0,0331	80
AMD	0,0191 (0,99)	0,0123	80
AMZN	0,0252 (0,98)	0,0122	80
BAC	0,0077 (0,53)	0,0036	80
FB	-0,0047 (-0,17)	0,0004	80
GOOG	0,0136 (0,75)	0,0072	80
GOOGL	0,0132 (0,67)	0,0058	80
MSFT	0,0093 (0,81)	0,0084	80
NFLX	-0,0077 (-0,37)	0,0018	80
NVDA	-0,0134 (-0,45)	0,0026	80

Table 9.2 Regression output for individual time-series regressions: dependent variable - residual stock returns, independent variable - weighed sentiment score

Ticker	Weighed Sent. Sc. Coeff.	Adjusted R-sq	# observations
AAPL	0,0008 (0,27)	0,001	80
AMD	0,0022 (0,45)	0,0026	80
AMZN	0,0045 (1,00)	0,0126	80
BAC	0,0027 (0,89)	0,001	80
FB	0,0056 (0,98)	0,0121	80
GOOG	0,0031 (0,89)	0,0101	80
GOOGL	0,0006 (0,14)	0,0003	80
MSFT	-0,0013 (-0,49)	0,003	80
NFLX	0,0049 (0,95)	0,0115	80
NVDA	-0,0041 (-0,74)	0,007	80

Table 10.1 Hausman test: dependent variable - stock returns, independent variable - normalized sentiment score

H_0: model is efficient under random effects

chi2(1)=3

p-value=0.0834

Table 10.2 Hausman test: dependent variable - stock returns, independent variable - weighed sentiment score

H_0: model is efficient under random effects

chi2(1)= 3.02

p-value=0.0825

Table 11.1 Likelihood ratio test: dependent variable - stock returns, independent variable - normalized sentiment score

H_0: homoscedasticity nested in heteroscedasticity

LR chi2(13)= -125.57

p-value=1.0000

Table 11.2 Likelihood ratio test: dependent variable - stock returns, independent variable - weighed sentiment score

H_0: homoscedasticity nested in heteroscedasticity

LR chi2(13)= -126.46

p-value=1.0000

Table 12.1 Wooldridge test for autocorrelation in panel data: dependent variable - stock returns, independent variable - normalized sentiment score

H_0: no first-order autocorrelation

LR chi2(13)= 206.921

p-value=0.0000

Table 12.2 Wooldridge test for autocorrelation in panel data: dependent variable - stock returns, independent variable - weighed sentiment score

H_0: no first-order autocorrelation

LR chi2(13)= 175.621

p-value=0.0000

Table 13.1 Panel data regression output: dependent variable - stock returns, independent variable - normalized sentiment score

Random-effects GLS regression				Number of obs	=	**1120**
Group variable: **TickerID**				Number of groups	=	**14**
R-sq: within	=	**0.0010**		Obs per group: min	=	**80**
between	=	**0.2940**		avg	=	**80.0**
overall	=	**0.0039**		max	=	**80**
				Wald chi2 (**1**)	=	**3.93**
corr(u_i, X)	=	**0** (assumed)		Prob > chi2	=	**0.0475**

(Std. Err. adjusted for **14** clusters in TickerID)

Stock return	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Sentnormalized	**0.0081594**	**0.0041178**	**1.98****	**0.048**	**0.0000887**	**0.0162301**
_cons	**-0.0005432**	**0.0007628**	**-0.71**	**0.476**	**-0.0020383**	**0.0009518**
sigma_u	**0.0014581**					
sigma_e	**0.01793397**					
rho	**0.00656689**	(fraction of variance due to u_i)				

Table 13.2 Panel data regression output: dependent variable - stock returns, independent variable - weighed sentiment score

Random-effects GLS regression				Number of obs	=	**1120**
Group variable: **TickerID**				Number of groups	=	**14**
R-sq: within	=	**0.0015**		Obs per group: min	=	**80**
between	=	**0.2419**		avg	=	**80.0**
overall	=	**0.0029**		max	=	**80**
				Wald chi2 (**1**)	=	**4.09**
corr(u_i, X)	=	**0** (assumed)		Prob > chi2	=	**0.0430**

(Std. Err. adjusted for **14** clusters in TickerID)

Stock return	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Sentweighed	**0.0016967**	**0.0008385**	**2.02****	**0.043**	**0.0000532**	**0.0033402**
_cons	**-0.0000862**	**0.0007**	**-0.12**	**0.902**	**-0.0014582**	**0.0012857**
sigma_u	**0.00160696**					
sigma_e	**0.01792933**					
rho	**0.00796906**	(fraction of variance due to u_i)				

* p<0.1, **p<0.05, ***p<0.01 : significance of z-statistic

Table 14.1 Robustness check on the full dataset. Panel data regression output: dependent variable - stock returns, independent variable - normalized sentiment score

Random-effects GLS regression				Number of obs	=	**23174**
Group variable: **TickerID**				Number of groups	=	**489**
R-sq: within = **0.0000**				Obs per group: min	=	**1**
between = **0.0011**				avg	=	**47.4**
overall = **0.0000**				max	=	**80**
				Wald chi2 (**1**)	=	**0.32**
corr(u_i, X) = **0** (assumed)				Prob > chi2	=	**0.5743**
			(Std. Err. adjusted for **489** clusters in TickerID)			
Stockreturn	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Sentnormalized	**0.0001499**	**0.0002669**	**0.56**	**0.574**	**-0.0003731**	**0.000673**
_cons	**-0.0015644**	**0.0001344**	**-12*** **	**0.000**	**-0.0018279**	**-0.001301**
sigma_u	**0.00115739**					
sigma_e	**0.01699914**					
rho	**0.00461422**	(fraction of variance due to u_i)				

Table 14.2 Robustness check on the full dataset. Panel data regression output: dependent variable - stock returns, independent variable - weighed sentiment score

Random-effects GLS regression				Number of obs	=	**23174**
Group variable: **TickerID**				Number of groups	=	**489**
R-sq: within = **0.0001**				Obs per group: min	=	**1**
between = **0.0000**				avg	=	**47.4**
overall = **0.0001**				max	=	**80**
				Wald chi2 (**1**)	=	**2.93**
corr(u_i, X) = **0** (assumed)				Prob > chi2	=	**0.0868**
			(Std. Err. adjusted for **489** clusters in TickerID)			
Stockreturn	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Sentnormalized	**0.0002792**	**0.000163**	**1.71***	**0.087**	**-0.0000404**	**0.0005987**
_cons	**-0.0015897**	**0.0001329**	**-12*** **	**0.000**	**-0.0018502**	**-0.001329**
sigma_u	**0.00116324**					
sigma_e	**0.0169984**					
rho	**0.00466114**	(fraction of variance due to u_i)				

* p<0.1, **p<0.05, ***p<0.01 : significance of z-statistic

Appendix 3: Machine learning analysis

Table 15. Machine learning stock returns prediction rates - unigram

Ticker	Naïve Bayes	Random Forest	SVM Gaussian	SVM Linear
AAPL	40%	44%	48%	40%
AMZN	**68%**	**64%**	**64%**	56%
FB	40%	**64%**	52%	56%
NVDA	48%	**68%**	**60%**	48%
GE	44%	44%	44%	**64%**
GOOGL	56%	48%	48%	40%
AMD	**72%**	**68%**	**68%**	**60%**
WMT	52%	**68%**	**68%**	52%
GOOG	40%	44%	44%	44%
BAC	40%	48%	**68%**	52%
T	56%	52%	48%	40%
MCD	40%	52%	28%	40%
MSFT	44%	36%	32%	44%
NFLX	56%	**64%**	**64%**	52%

Table 16. Machine learning stock returns prediction rates - bigram

Ticker	Naïve Bayes	Random Forest	SVM Gaussian	SVM Linear
AAPL2	-	48%	48%	40%
AMZN2	-	**64%**	**64%**	56%
FB2	-	56%	52%	64%
NVDA2	-	**60%**	**60%**	44%
GE2	-	44%	44%	60%
GOOGL2	-	48%	48%	40%
AMD2	-	**68%**	**68%**	**64%**
WMT2	-	**68%**	**68%**	**60%**
GOOG2	-	44%	44%	40%
BAC2	-	40%	68%	44%
T2	-	48%	48%	44%
MCD2	-	**60%**	28%	**60%**
MSFT2	-	32%	32%	44%
NFLX2	-	**64%**	**64%**	**60%**

Table 17. Machine learning stock returns prediction rates - trigram

Ticker	Naïve Bayes	Random Forest	SVM Gaussian	SVM Linear
AAPL3	-	48%	48%	44%
AMZN3	-	**64%**	**64%**	**64%**
FB3	-	52%	52%	**68%**
NVDA3	-	**60%**	**60%**	52%
GE3	-	44%	44%	**64%**
GOOGL3	-	48%	48%	40%
AMD3	-	**68%**	**68%**	**60%**
WMT3	-	**68%**	**68%**	56%
GOOG3	-	44%	44%	44%
BAC3	-	32%	**68%**	56%
T3	-	48%	48%	56%
MCD3	-	36%	28%	64%
MSFT3	-	32%	32%	40%
NFLX3	-	**64%**	**64%**	**64%**

Declaration of Authorship

We, Levi Bergstrand and Maksim Khritantsev, hereby certify that this thesis has been composed by us and is based on our own original work unless stated otherwise. No other person's work has been used without due acknowledgment in this thesis. All references and verbatim extracts have been quoted, and all sources of information have been specifically acknowledged.

This dissertation was not previously presented to another examination board and has not been published.

LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

The authors can be reached at their personal e-mails

Authors:

Levi Bergstrand - bergstrand.levi@gmail.com

Maksim Khritantsev - maxim.khritantsev@gmail.com

The supervisor can be reached at her university e-mail

Supervisor:

Olga Balakina - olga.balakina@nek.lu.se

Lund, 2018