universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Social Media and its relationship to the stock market – Correlation of Tweets and the EURO STOXX 50“

verfasst von / submitted by

Ramona Gassner, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 915
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Betriebswirtschaft UG2002
Betreut von / Supervisor:	a.o. Univ.-Prof. Mag. Dr. Christian Keber

I. Abstract

English

This research discusses the question whether the information about the European economy posted on Twitter is related to the European stock market. The assumption is that the market does not distinguish between positive or negative information. Thus, the investigation focuses on a correlation analysis of the number of Tweets posted within a given time interval and the EURO STOXX 50, representing the European market. Further analysis focuses on the exploration of a forecasting model. With a linear regression, the author predicted intraday prices of the EURO STOXX 50, based on historical EURO STOXX 50 prices and the number of Tweets. It has been found that the number of Tweets posted within a time interval of one minute and the respective intraday price of the EURO STOXX 50 has a significant weak negative correlation. This result shows that when the number of Tweets is high, the variable "EURO STOXX 50" is rather low. Further, the forecast model predicted a forecast accuracy of the movement of the EURO STOXX 50 lying higher than 50%.

German

Die vorliegende Arbeit befasst sich mit der Fragestellung, ob Twitter Meldungen über Informationen der europäischen Wirtschaft, einen Zusammenhang mit der Entwicklung des europäischen Aktienmarktes darstellen. Es liegt die Annahme zugrunde, dass der Markt nicht zwischen positiv oder negativ formulierten Meldungen unterscheidet. Die Studie untersucht die Korrelation zwischen der Anzahl an geposteten Tweets in einem bestimmten Zeitintervall und den EURO STOXX 50 Preisen pro Minute. Im weiteren Teil der Arbeit wird ein EURO STOXX 50 Prognosemodell erstellt. Mithilfe einer Einfachregression errechnet der Autor EURO STOXX 50 Preise, basierend auf historischen Index-Preisen und der Anzahl an Tweets. Zwischen der Anzahl an geposteten Tweets pro Minute und dem EURO STOXX 50 Preis pro Minute besteht eine signifikant schwache Negativkorrelation. Dieses Ergebnis zeigt, dass, wenn die Anzahl der Tweets hoch ist, der EURO STOXX 50 Preis eher niedrig ist. Das Prognosemodell, das die Veränderungen des EURO STOXX 50 Preises zeigt, hat eine Prognosegenauigkeit von mehr als 50 %.

II. Acknowledgement

I would like to thank all the involved persons, who have supported me throughout different phases of writing this thesis. I am sincerely thankful to my sister Caro and my boyfriend Martin, who never got tired of discussing with me the methodology and statistics used in this thesis. I express my gratitude to my managers who gave me the necessary support for writing this thesis, even during tough project periods. Further, I would like to acknowledge my parents for giving me the opportunity to reach the point of finishing my university degree. Finally, I thank my professor, who guided and supported me until the very end.

III. Table of contents

IV. List of figures

V. List of tables

1 Introduction and background

1.1 Introduction

The internet offers researchers and analysists the possibility to explore and investigate a massive quantity of data–so-called “big data”–and allows them to easily extrapolate meaningful information out of an unstructured volume of spread news and comments. An article in the German newspaper Die Zeit describes a phenomenon according to which the Berkshire Hathaway stock increases after the release of an Anne Hathaway movie. As described in the research of an American blogger–who calls this phenomenon the “Hathaway-effect”–this circumstance is due to text analysis algorithms, which search for specific words on social media and match the word “Hathaway” automatically to the respective company Berkshire Hathaway. This brings investors to the assumption that market participants have a high interest in the stock, which leads to an increase in the price (Crocoll, 2012). Another phenomenon, which also belongs in this category, is the case of the American rapper 50 Cent, who promoted the American penny stock “H&H Imports” on his Twitter account. After mentioning this on Twitter, the penny stock increased more than 300% (Hoyer, 2011). Social media also plays a big role when it comes to rumors related to the economy and financial markets. On January 22, 2013, the German index DAX lost within a very few minutes more than 100 points. It is possible that this was due to a Twitter mention, which stated that the president of the German Central Bank had resigned (Goldberg, 2014). Obviously, this was incorrect information, but the market reacted quickly. The question arises whether even incorrect information spread on social media can influence the stock prices significantly. It is important to note that news and information is spread very quickly without any verification of the information The phenomena mentioned above immediately suggest that there might be a relationship or even an influence between Twitter data–independent of what kind of information is spread–and the stock market.

This investigation handles the question of whether or not there is a statistically significant relationship between information spread on social media and the stock market. Furthermore, presuming that there is a significant relationship between Tweets and the index price, the study investigates if the number of Tweets influences the EURO STOXX.

Hence, the project investigates whether the collected data set allows a forecasting model that can predict accurate prices for the EURO STOXX 50.

The research paper starts with an introduction to the theoretical background of the financial market hypothesis, and explains how models predicting the future values of stocks are determined. Chapter 1 continues with an explanation of behavioral finance theories. Detailed definitions concerning social media, social networks, Twitter and big data can be found in Chapter 1.3. The chapter closes with an introduction to sentiment analysis.

Chapter 2 describes the field of investigation and explains the hypotheses treated within this research paper. Research and statistical methodology is described in Chapter 3. The research results and the discussion close the paper with Chapters 4 and 5.

1.2 Financial market theories

A widely accepted theory regarding financial markets claims that predicting the price of a security is not possible. Researchers disagree as to whether this statement holds or can be falsified. The following section deals with capital market theory and discusses (1) the efficient market hypothesis and (2) the random walk hypothesis before explaining behavioral finance theories. Behavioral finance theories highlight inefficiencies in the financial markets, such as under- or over-reactions to information.

1.2.1 The efficient market hypothesis

The efficient market hypothesis says that prices fully reflect all available information about a security. When new information is available about the market, it will immediately correct the value of the securities (Fama E., 1965). Thus, it is impossible for investors to buy undervalued stocks or sell overvalued stock. The precondition for this hypothesis belongs to Grossman and Stiglitz (1981) who state that information and trading costs are always zero. Jensen (1978), on the other hand, says that prices reflect information up to the point where the marginal benefit of information–the profit of price differences gained by a potential information gap–do not exceed the marginal costs.

Fama (1970) divides his work on market efficiency into three main questions: (1) How well do past returns predict future returns? (2) How quickly do security prices reflect public information announcements? (3) Do any investors have private information that is not fully reflected in market prices? The answer to the first two questions can be seen in event studies. Event studies on daily returns give evidence that market efficiency holds when an information event can be dated precisely and the event has a large effect on prices. These studies give a picture of the speed at which prices adjust to new information. The literature indicates that on average, stock prices adjust quickly to information about investment decisions, dividend changes, changes in capital structure, and corporate-control transactions. This outcome proves that prices adjust efficiently to firm-specific information (Fama E., 1991).

1.2.2 The random walk hypothesis

Malkiel (1973) claims that the price movement of a stock is no more predictable than the random selection of successive steps in the positive, negative or neutral direction of the value of the stock. The efficient market theory goes together with a "random walk," which characterizes a price series in which all subsequent price changes represent random changes from previous prices.

The random walk theory says that someone cannot predict future steps and that directions cannot be predicted based on past actions. This means–following the efficient market theory–that a stock price incorporates all the available information about the value of the stock and that short-term changes in the stock prices cannot be predicted. This means that if all assets in a market are correctly priced, nobody would gain or lose by trading (Malkiel B. G., 1973).

1.2.3 Technical and fundamental analysis of the market

Technical and fundamental analysts take the opposite opinion: they determine psychological and behavioral elements of stock prices, and have come to believe that future stock prices are somewhat predictable based on past stock price. Technical analysis is the analysis of past stock prices to predict future prices. Fundamental analysis, on the other hand, is the analysis of financial information such as company earnings, asset values, etc., to help investors select "undervalued" or "overvalued" stocks (Malkiel B. , 1989). Technical analysts believe that the market is 10% logical and 90% psychological. However, fundamental analysts believe that the market is 90% logical and only 10% psychological. Both have in common that they argue that they can predict future prices by analyzing past prices (Malkiel B. G., 1973).

1.2.4 Behavioral finance

Behavioral finance theories follow fundamental analysis, arguing that financial decisions are psychological rather than rational (Tversky & Kahneman, 1974).

This theory argues that investors are biased. They are overconfident and overoptimistic because they overestimate their ability and the accuracy of the information they have. Further, they assess situations based on superficial characteristics rather than underlying

probabilities. Investors are conservative in the sense that their forecasts have a bias towards prior beliefs over new information (Byrne & Brooks, 2008).

Considering the examples we mentioned in the introduction–e.g., that investors rely on information which is not verified, like rumors–this research project assumes that many investors act in a very overconfident and overoptimistic way with information they receive. If the market were efficient, all the incorrect information would be reflected in the price. This might lead to the assumption that there is also information reflected in the current price, which is not related to the value of the stock. Consequently, this leads to a mispricing. Furthermore, if there is a mispricing in the security, it would be possible to predict future prices by analyzing the past information and detecting patterns, correlations or discrepancies in the prices.

Before explaining the field of investigation, the following sections first describe social media and social networks and explain on a deeper level the social media channel Twitter, and the opportunity Twitter provides of exploring the financial market.

1.3 Social media and social networks

In 2008, Boyd and Ellison defined social network sites as any web-based service that allows individuals to have a (semi-) public profile and to have the possibility of articulating statements and social data which they share to other users within a bounded system. Due to the interconnected, social, rapid, and public exchange of information, we speak about "information that can go viral." This phrase was used in early 2004 for the first time. The phenomenon of always available, viral information changed the way we communicate with each other. Today's communication is rather bi-directional and "many-to-many" than isotropic and "one-to-many." Isotropic and "one-to-many" describe old-fashioned social media like radio broadcasters and television (Danneman & Heimann, 2014).

In 2014, Danneman and Heimann described social data as data in textual form produced by people for other people's consumption.

Makhaber (2015) defines the essential characteristics of a social network as follows:

- A collection of entities that participate in a network (typically these entities are people)

- At least one relationship between the entities of the network exists (e.g., on Facebook, this relationship is called “friends”)
- An assumption of non-randomness or locality is given (e.g., relationships tend to cluster: A is related to B and C, etc.)

Given this definition, telephone networks, e-mail networks and collaboration networks also fall under this social network definition.

1.3.1 Twitter

Due to its real-time nature and massive access to big data, the microblogging service Twitter’s fast growth within the last years has drawn much attention from researchers (Mao, 2012). Twitter was launched March 21, 2006, and now sends one billion Tweets every 2.5 days. There is one specific difference to Twitter as compared with other blogs or consumer reviews: Twitter limits its users to a document length of 140 characters–therefore it is called “microblogging.” This brevity makes sentiment analysis much simpler because users tend to be pithy and accurate rather than loquacious and artful (Danneman & Heimann, 2014).

Big data is, on the one hand, a meaningful source of information, but on the other hand, there is the fact that researchers must mine for relevant topics while ignoring the noise and spam that surround them (Wolfram, 2010). Twitter messages may allow researchers to limit noise due to its accuracy in the lengths of the messages.

Furthermore, Twitter enables access to millions of user opinions, from a variety of users from different backgrounds and professions (Wolfram, 2010). A further important common characteristic of microblogs like Twitter is its real-time nature. A Twitter user updates their blog several times a single day. It is to some extent possible to check and get information about what users are doing and thinking at any time. Users also post social events and include disastrous events such as storms, fire, traffic jams, etc., and thus Twitter is also helpful for various real-time necessities such as helping during a fire emergency or providing traffic updates. For example, it is possible to detect an earthquake occurrence promptly when observing the Tweets people post related to the earthquake (Sakaki, Okazaki, & Matsuo, 2010).

1.3.2 Big data

People create nearly 2.3 million terabytes (2.5 quintillion bytes) of data every day, indeed 90% of the data in the world today has been created in the last two years alone. Big data is both; it is the deluge of data being generated and the astronomical size of data sets themselves. The role of data scientists is to create challenges and opportunities in new searching fields for both factors (Danneman & Heimann, 2014). Makhabel (2015) describes big data with three major characteristics: volume, variety and velocity. Velocity defines the data process rate, or how fast the data are being processed. Variety denotes the data source types and volume describes the size of the data. The table from Makhabel (2015) shows in a very illustrative way how data size has grown within the last 30 years.

Year	Data Sizes	Comments
N/A		1 MB (Megabyte) = 2^{20}. The human brain holds about 200 MB of information.
N/A		1 PB (Petabyte) = 2^{50}. It is similar to the size of 3 years' observation data for Earth by NASA and is equivalent of 70.8 times the books in America's Library of Congress.
1999	1 EB	1 EB (Exabyte) = 2^{60}. The world produced 1.5 EB of unique information.
2007	281 EB	The world produced about 281 Exabyte of unique information.
2011	1.8 ZB	1 ZB (Zetabyte)= 2^{70}. This is all data gathered by human beings in 2011.
Very soon		1 YB(Yottabytes)= 2^{80}.

Table 1 "Data sizes" (Makhabel, 2015, S. 9)

The aim for researchers executing data mining and programming-specific algorithms is to perform in real time. Therefore, the Internet allows us to connect directly and work with big data contemporarily.

Some companies consider 10 TB (terabyte; 1 TB = 10^{12} Byte = 1 000 000 000 000 Byte) to be big data while others regard 1 PB (petabyte; 1 PB = 1000 TB) as the threshold for big data. Some of the popular social media networks that hold big data are as follows (Prajapati, 2013):

- Facebook has 40 PB of data and captures 100 TB/day

- Yahoo! has 60 PB of data
- Twitter captures 8 TB/day
- eBay has 40 PB of data and captures 50 TB/day

1.3.3 Sentiment analysis

The field of analyzing people's opinions, sentiments, evaluations, attitudes, and emotions through written language is called sentiment analysis, opinion mining or data mining. In the past, opinion mining hardly existed. This is because opinions were elicited in surveys rather than in text documents. Within the last years, social data mining has become more and more important due to the explosion of sentiment-laden content on the Internet (Danneman & Heimann, 2014).

Danneman and Heimann (2014) describe data mining as a set of tools and techniques used to describe and make inferences about data. Makhabel (2015) also includes algorithmic problem solving in the definition of data mining. He states that clustering, classification, association rule learning, anomaly detection, regression, and summarization are all part of the tasks belonging to data mining. Hence, the data mining methods can be summarized into two main categories of problems: summarization and feature extraction. According to Makhabel (2015), the target of summarization is to cluster and examine a collection of points (data) and group the points according to some measures. The goal for this category is to create points in the same cluster which have a small distance from one another, while points in different clusters are at a large distance from each other. On the other side, there is the category of feature extraction, which describes an extraction of the most prominent features of the data and ignores the rest. For example, all the data consist of baskets of small sets of items, or the data look like a collection of sets and the objective is to find pairs of sets that have a relatively large fraction of their elements in common (Makhabel, 2015). For this thesis, the category of feature extraction is used. This means similar data items (Twitter messages) are collected into a basket (the financial market EURO STOXX 50) and the aim is to find dependencies respectively with the stock market.

1.4 Recent studies on the correlation of Tweets and the stock market

The first empirical study dealing with internet stock message boards and financial markets was Wysocki in 1999. He reports that message postings for 50 firms did forecast the next trading day trading volume and the next day abnormal stock returns. Furthermore, he found that the firms with high message-posting activity were characterized by high market valuation relative to fundamentals, high short-seller activity, high trading volume and high analyst following, but low institutional holdings (Wysocki, 1991).

In 2009, Schumaker and Chen tried to forecast the actual price of S&P 500 listed stocks using a Support Vector Regression (SVR) algorithm underlying text mining techniques of financial news articles. They found a forecast model that predicts future prices best when taking text data from the release of the news articles.

In 2012, Mao et al. found that the daily number of Tweets is correlated with the price of the S&P 500 at the aggregated index level, but also when dividing the index into industry sectors or even individual company stocks. Additionally, they figured out that Twitter data helped to predict the stock market.

Although many previous studies show us how hard it is to predict stock returns, Antweiler and Murray see it is as plausible that social media messages (they used the platforms Yahoo! Finance and Raging Bull for their study) may help to predict future returns. They argue that a very high proportion of the messages in social media channels contain explicit hints regarding particular stocks and the buy, sell, hold question. This means that stock messages reflect public information extremely rapidly, which means that there is financially relevant information present. They used the Naïve Bayes algorithm and the Support Vector Machine to interpret text messages, and found that there is useful information present on the stock message boards. In their test, the message boards do not successfully predict stock return, while they show that message posting helps to predict subsequent trading volume and volatility (Antweiler & Murray, 2005).

The main finding of Wolfram's study in 2010 was that social media posts (in this case the microblog Twitter) can be used to build a close model of stock price predictions which contradicts the efficient market hypothesis (Wolfram, 2010).

Another study from Tayal and Komaragiri in 2009 focused on sentiment analysis of blogs and micro-blogs to uncover their predictive power on stock prices. Their experimental result was to predict the actual stock price of the following day from the models of a social media data source (Tayal & Komaragiri, 2009).

2 Field of investigation and hypotheses

2.1 Data mining of European Twitter data

Twitter data as an informational and data mining source for researchers have been used for several studies, mainly based on American data, e.g., S&P 500 companies (Wolfram, 2010 or Mao, 2012). These studies have in common a very high sample size (e.g., number of Tweets for a given time interval), because there is a lot of US-relevant information shared on Twitter. This study treats the question whether it is possible to collect a sample of European Twitter data, with a data set of European companies included in the EURO STOXX 50 index, allowing statistical research for data mining.

Hypothesis one states that company-relevant information about the 50 largest listed European companies (based on the companies included in the EURO STOXX 50 index as of September 12, 2016) has enough mentions on Twitter such that statistically meaningful data mining can be explored.

Another interesting question is to see whether the data show special events relevant to the economy or financial markets. These special events change the usual tweeting habit of users and lead to an increase in the spread of information about various EURO STOXX companies; hence, we investigate if the number of Tweets is noticeably higher.

2.2 Correlation of number of Tweets and index indicators

Besides the investigation of the data set, we want to explore whether Tweets including information on the EURO STOXX 50 companies are related to the movement of the index price. We want to see if there is a significant correlation between the number of Tweets and the EURO STOXX 50.

Hypothesis two asks if the number of Tweets and the EURO STOXX 50 prices have a significant correlation based on observed data for a one-minute time interval.

Furthermore, we will explore whether the changes in the price of the EURO STOXX 50 (one-minute time interval) are correlated to the number of Tweets on an absolute and relative level. The price change of the EURO STOXX 50 may indicate that a special event within the economy or financial market occurred which significantly increased or decreased the index price.

Additionally, we want to see the price and Twitter movements in time and ask if there is a correlation visible over time.

2.3 Forecasting model

Hypothesis three asks if it is possible to find a forecasting model which can predict the EURO STOXX 50 prices. Specifically, we want to test whether the model used by Mao (2012) fits our data, and if it would enable us to predict future values. Mao (2012) applied a linear regression with an exogenous input model to Twitter predictors and the stock market indicators. Mao found that including Twitter data is helpful in building more accurate prediction models. The prediction accuracy was 68%, which is significantly more accurate than random guessing. In this chapter, we compute the prediction accuracy of our data model.

3 Methodology

Addressing the questions and hypotheses asked in the section "Field of investigation and hypotheses," we illustrate the methodology and approach in the following section. This section starts with an introduction concerning the data collection methodology and ends with a walkthrough of the statistical methodology and employed concepts.

3.1 Data collection and preparation

The data collection has been conducted for two types of variables. On the one hand, we collected a financial data set of the EURO STOXX 50 index prices, and on the other hand we collected a data set of the number of Tweets mentioning the listed companies of the EURO STOXX 50 index.

3.1.1 EURO STOXX 50 data

Since the information on social media in general, and specifically on Twitter, is spread very rapidly (big data), we collected EURO STOXX 50 data on a one-minute intraday interval. The data collection has been conducted with registered access to and regular downloads of index prices from Thomson Reuters.[1] The observational period took place from September 13 to October 19, 2016. Due to the calculation of the EURO STOXX 50 index between 09:00 (09:00 am) and 17:30 (05:30 pm) the data set of the financial data is limited to the given time period (STOXX Limited, 2016).

3.1.1.1 Data cleaning of financial data

According to Makhabel (2015) data cleaning is the task which helps to fill missing values, smooth out noise and correct inconsistencies in the data. The aim of data cleaning is to improve the data quality of the data set with respect to the following considerations:

- Accuracy: Are the data recorded correctly?
- Completeness: Are all relevant data recorded?
- Uniqueness: Are there no duplicated data recorded?

[1] https://www.thomsonone.com/

- Timeliness: How old are the data?
- Consistency: Are the data coherent?

In general, the process of data mining contains two steps. The first step is to find discrepancies in the source of the data set. The second step is to correct discrepancies within the transformation into a statistics program (Makhabel, 2015). In the following, we will discuss the above-mentioned points based on these two steps, and describe the conducted workarounds:

Accuracy: In order to detect incorrect data, we cross-checked the opening and closing prices of the EURO STOXX 50 data with other sources (e.g., Yahoo! Finance, Google). We noticed that for our data set from Thomson Reuters, the opening and closing prices for September 22, 2016 did not correspond with other sources. Due to this fact, we removed all data from September 22, 2016 from the data set. To avoid a data gap in the statistics program SPSS, we defined the value cell as a numerically coded value (a dummy value) to represent the missing data point. This value tells the statistic program that there is no recorded value and the program ignores the appropriate data cell (Field, 2005).

Completeness: For October 7, 2016, there were no data collected. To avoid a data gap in the statistics program, we defined the missing value cells as a numerically coded dummy value.

Uniqueness: In the financial data set there are no duplicated data. Each financial data point corresponds to the respective date (DD.MM.YYYY) and respective time (hh:mm).

Timeliness: The observational period and the collection of the financial data have been conducted at the same time as the collection of the Twitter data. There are no time discrepancies within the data.

Consistency: Since the financial data have been collected with a one-minute time interval, the consistency of the data is ensured.

3.1.2 Twitter data

Twitter offers an open source application programming interface (API), which allows researchers to collect Twitter data on a real-time basis. It is important to note that it is not possible to download historical Twitter data. This means that for our data collection, a stable internet connection was essential. The Twitter API can relate to various general-purpose

programming software. For this research, we used the statistical computing environment "R." The data collection was conducted from September 13 to October 19, 2016 between 09:00 (09:00 am) and 17:30 (05:30 pm). This observational period corresponds with the intraday calculation time and data collection of the EURO STOXX 50 index. After the data collection with "R," we transferred the outcome (number of Tweets, content of Tweets and corresponding day and time) to Excel and SPSS.

3.1.2.1 Twitter API and data collection with "R"

The collection of Twitter data ran with a developer account based on a registered user account in the Twitter environment. Within the set-up of the Twitter developer account, the user interface asks for a personal homepage and certain descriptions of the objective of the account. After completion of the developer account, the API can relate to the programming software "R." Before doing this, the Twitter environment for "R" needs to be installed (see Appendix for program codes for installation of Twitter libraries). To get a connection to the real-time interface an authentication process is needed. This happens with the so-called "OAuth settings" (see Figure 1) (Danneman & Heimann, 2014).

```
Authentithication:

credential <- OAuthFactory$new(consumerKey='[redacted]',
                               consumerSecret='[redacted]',
                               requestURL='https://api.twitter.com/oauth/request_token',
                               accessURL='https://api.twitter.com/oauth/access_token',
                               authURL='https://api.twitter.com/oauth/authorize')

options(RCurlOptions = list(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl")))
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

credential$handshake(cainfo="cacert.pem")
```

Figure 1 OAuth setting in "R"

3.1.2.2 Keyword description

When "R" is connected to the Twitter API and the user is authenticated, the request for Twitter data can be started. The function "*FilterStream*" opens a connection to the API that returns Tweets which match one or more filter predicates. All published Tweets can be filtered by keywords, users, language and location. Afterwards, the output can be saved (Barbera, 2014).

The filter predicates for this research correspond to keywords of the 50 companies listed in the EURO STOXX 50 stock index. The table of the listed companies as of September 2016 can be found in the Appendix (Table 15).

It lists 50 of Europe's blue chip companies from 12 Eurozone countries (Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain) covered by the index as of September 12, 2016 (STOXX Limited, 2016).

The filter request contains the stock market ticker symbol of each of the components of the EURO STOXX 50 index and also contains the company's full firm name. This method guarantees that all relevant data are recorded. Besides the companies' names, the request asks also for keywords like "EuroStoxx" or the ticker symbol of the EuroStoxx50 "SX5E" (see Figure 2 API request for EURO STOXX 50 Tweets in "R").

```
filterStream( file.name="tweets_eurostoxx_0709.json",
              track= "$ABI.BR, $AI.PA, $AIR.PA, $ALV.DE, $ASML.AS, $BAS.DE, $BAYN.DE, $BBVA.MC,
$BMW.DE, $BN.PA, $BNP.PA, $CA.PA, $CS.PA, $DAI.DE, $DBK.DE,$DG.PA, $DPW.DE, $DTE.DE, $EI.PA,
$ENEL.MI, $ENI.MI, $EOAN.DE, $FP.PA, $G.MI, $GLE.PA, $GSZ.PA, $IBE.MC, $INGA.AS, $ISP.MI, $ITX.MC,
$MC.PA, $MUV2.DE, $NOK1V.HE, $OR.PA, $ORA.PA, $PHIA.AS, $REP.MC, $RWE.DE, $SAN.MC, $SAN.PA, $SAP.DE,
$SGO.PA, $SIE.DE, $SU.PA, $TEF.MC, $UCG.MI, $UL.PA, $UNA.AS, $VIV.PA, $VOW3.DE, ABI.BR,Anheuser-Busch
InBev SA/NV, AI.PA, L'Air Liquide, AIR.PA, AIRBUS GROUP, ALV.DE, Allianz, ASML.AS, ASML HLDG, BAS.DE,
BASF, BAYN.DE, Bayer, BBVA.MC, BBVA, BMW.DE, Bayerische Motoren Werke, BN.PA, Danone, BNP.PA, BNP
Paribas, CA.PA, Carrefour, CS.PA, AXA, DAI.DE, Daimler, DBK.DE, DeutscheBank, DG.PA, VINCI, DPW.DE,
DeutschePost, DTE.DE, Deutsche Telekom, EI.PA, Essilor International, ENEL.MI, Enel, ENI.MI, Eni,
EOAN.DE, E.ON, FP.PA, TOTAL, G.MI, Assicurazioni Generali, GLE.PA, Societe Generale, GSZ.PA, ENGIE,
IBE.MC, IBERDROLA, INGA.AS, ING GROUP, ISP.MI, Intesa Sanpaolo, ITX.MC, INDITEX, MC.PA, LVMH Moët
Hennessy Louis Vuitton, MUV2.DE, Münchener Rückversicherungs-Gesellschaft, NOK1V.HE, Nokia Corporation,
OR.PA, L'Oreal, ORA.PA, PHIA.AS, ROY.PHILIPS, REP.MC, REPSOL, RWE.DE, RWE, SAN.MC, BANCO SANTANDER,
SAN.PA, Sanofi, SAP.DE, SAP, SGO.PA, Compagnie de Saint-Gobain, SIE.DE, Siemens, SU.PA, Schneider
Electric, TEF.MC, TELEFONICA, UCG.MI, UniCredit, UL.PA, UNIBAIL-RODAMCO, UNA.AS, UNILEVER CERT, VIV.PA,
Vivendi, VOW3.DE, Volkswagen, EUROSTOXX, EUROSTOXX50, $SX5E, SX5E, ^STOXX50E, STOXX", oauth=credential,
timeout=42400, verbose=TRUE)
```

Figure 2 API request for EURO STOXX 50 Tweets in "R"

3.1.2.3 Data cleaning and data preparation

After running the filter request, we cleaned our data to remove noise and correct inconsistencies or data gaps:

- **Accuracy**: To have accurate data in the sample, it is necessary to be sure that the keywords in the request do not have a double meaning. In order to avoid this noise, generic words and abbreviations like "Aktiengesellschaft," "AG," "SA," "S.p.A," "bank," "post" and "orange" have been removed from the request beforehand.

- **Completeness**: Since the request is very much dependent on a stable internet connection, gaps do appear in the data. This is a mechanical fault and leads to incomplete data. However, for statistical computing we chose a numerically coded dummy value to represent the missing data point. This value tells the statistics program that there is no recorded value, and the program ignores the respective data cell (Field, 2005).

- **Uniqueness**: In the Twitter data set there are no duplicated data. Each data value (Tweet) corresponds to the respective date, respective time respective user and location.

- **Timeliness**: The observational period and collection of the Twitter data have been conducted at the same time as the collection of the financial data. There are no time discrepancies within the data.
- **Consistency**: Since the Twitter data have been collected with a constant programming interface, the consistency of the data is ensured.

The output of the "R" request was saved with a JSON file. To get a better overview of the content, we transferred the data into Excel. Each Tweet contains specific information (see Figure 3) about the following:

- date and time of creation,
- Twitter ID number,
- text of the Tweet with posted hyperlinks,
- source of the Tweet,
- information about Tweet creation (e.g., retweet, responding to an existing Tweet, newly created Tweet),
- name of the user and the number of his or her followers,
- the language of the Tweet
- information about certain internal default settings.

{"created_at":"Thu Oct 06 12:55:16 +0000 2016"	id:784014161298661376	id_str:"784014161298661376"	text:"News: Repsol S.A. (REPYY: OTCQX International Premier) https:\/\/t.co\/iEGCffW015"
{"created_at":"Thu Oct 06 12:55:18 +0000 2016"	id:784014168986902528	id_str:"784014168986902528"	text:"RT @PAHCBages: Marxem de #BagesVsBBVA amb comprom\u00eds per part del @bbva d
{"created_at":"Thu Oct 06 12:55:18 +0000 2016"	id:784014169561559041	id_str:"784014169561559041"	text:"Par for the western pharma-agri science course. Loot the 1st world buyer. Then loot the 3
{"created_at":"Thu Oct 06 12:55:19 +0000 2016"	id:784014172883394560	id_str:"784014172883394560"	text:"Caralho
{"created_at":"Thu Oct 06 12:55:21 +0000 2016"	id:784014178260574208	id_str:"784014178260574208"	text:"#Bayer completes sale of Bayer Garden and Bayer Advanced businesses to SBM #crops htt
{"created_at":"Thu Oct 06 12:55:21 +0000 2016"	id:784014178197667840	id_str:"784014178197667840"	text:"Carlos Torres Vila en South Summit: \"BBVA quiere ser m\u00e1s que un banco
{"created_at":"Thu Oct 06 12:55:21 +0000 2016"	id:784014180655431680	id_str:"784014180655431680"	text:"RT @PAH_BCN: Felicitats comPAHnys #SiSePuede https:\/\/t.co\/qbGhTdB3Wu"
{"created_at":"Thu Oct 06 12:55:21 +0000 2016"	id:784014181272055808	id_str:"784014181272055808"	text:"I migliori amici - Scarica l'App e Vinci un Week-end con i #BR3 ! #PIUMA https:\/\/t.co\/hE
{"created_at":"Thu Oct 06 12:55:23 +0000 2016"	id:784014187659825152	id_str:"784014187659825152"	text:"@eczozgurozel eni\u015fte komisyonu kurulmul 15 dk ile darbeyi \u00e7\u00f6zerler"
{"created_at":"Thu Oct 06 12:55:23 +0000 2016"	id:784014188922503168	id_str:"784014188922503168"	text:"RT @LA_PAH: \u00bfVeis como S\u00cd SE PUEDE y que somos imPAHrables? \u00a1ENH

Figure 3 Output of real-time Twitter data

For dealing further with the data, we prepared the Twitter output for two different approaches. In one approach, we prepared the text of the Tweets with the respective date and time for certain text sentiment analysis (Excel and JSON file). In the other approach, we summed up the number of Tweets per day and per minute, to further investigate the variable “number of Tweets per day/per minute.”

3.2 Statistical methodology and approach

The statistical analysis was conducted with SPSS and the Data Analysis add-in in Excel for cross-checking outcomes. Statistical graphs have been created with Excel. Statistical citations, methods and models, if not stated otherwise, correspond to Andy Field’s “Discovering Statistics Using SPSS” (2005).

3.2.1 Data mining of European Twitter data

Hypothesis one asks whether company-relevant information about the 50 largest listed European companies has enough mentions on Twitter so that statistically meaningful results can be explored. We conducted a time series analysis of the variable “number of Tweets” and evaluated with a descriptive statistical analysis the sample and its distribution. The test statistics show us if the data set contains statistical significant results.

3.2.1.1 Frequency of Tweets per day and text analysis

A time series table of the variable “number of Tweets” gives us an understanding of the data set within the observation time. We wanted to figure out how many Tweets we collected during a day. Furthermore, we analyzed the given time series plot to see if the Tweets occurred largely in the morning or in the afternoon and highlight outliers in the

data. With a text sentiment analysis of the Tweets we investigated which information triggered the highest number of Tweets.

Outliers are scores that are very different from the rest of the data and bias the data model, but play a major role for this research. These outliers are special events within the investigational period which led Twitter users to change their usual habits of sharing information and to increase the number of Tweets, hence the amount of spread information was unusually high. We defined outliers as data points which are more than twice and more than three times the given standard deviation:

$$Outlier_1 = 2sd$$

$$Outlier_2 = 3sd$$

The time series plot shows us these certain outliers. To look at the outliers, we standardized the data set ($z_{(xi)}$) to easily define benchmarks (Field, 2005):

$$Z_{(x_i)} = \frac{x_i - \bar{x}}{sd}$$

sd … standard deviation
$\bar{x}$ … $mean$

As soon as these outliers were identified we conducted a text sentiment analysis to find out what information on the economy and financial markets triggered this high number of Tweets.

The text analysis contains the Tweets of the certain outlier days. We sum up the most frequent words in the Tweets with the program code in Figure 4 (Anderson, 2017) and explore the content of these Tweets.

```
temple.text<-scan(choose.files(),what="char",sep="\n")
temple.text<-tolower(temple.text)
temple.words.list<-strsplit(temple.text,"\\W+",perl=TRUE)
temple.words.vector<-unlist(temple.words.list)
temple.freq.list<-table(temple.words.vector)
temple.sorted.freq.list<-sort(temple.freq.list,decreasing=TRUE)
temple.sorted.table<-paste(names(temple.sorted.freq.list),temple.sorted.freq.list,sep="\t")
cat("Word\tFREQ",temple.sorted.table,file=choose.files(),sep="\n")
```

Figure 4 Program code for text sentiment analysis

In this way, we gathered keywords of the most frequent words and searched them on the Internet to get an understanding of what news was spread around the globe.

3.2.1.2 Variable number of Tweets per minute

For further investigation, we concentrated on the number of Tweets within minute-to-minute intervals, meaning, we summarized the number of Tweets published within a minute. The outcome is a frequency table underlying the time between the opening and closing hours of the European stock market, from 09:00 am to 05:30 pm. A time series plot shows us the movement of the variable "time" on the x-axis and the number of Tweets published within a minute on the y-axis. The box-and-whisker chart shows us graphically the minimum, maximum, quantiles, median and the outliers of the variable.

3.2.1.2.1 Descriptive statistics: univariate analysis of the variable "number of Tweets per minute"

The frequency distribution, its skewness and kurtosis of the variable "number of Tweets per minute" can be indicated with a histogram. The univariate analysis shows the mean, standard deviation and variance, median, mode, minimum and maximum. If the skewness variable is positive, the distribution is positively skewed and the number of Tweets is clustered at the lower end (e.g., the number of Tweets per minute is more frequent between zero and the lower scores). When this is the case, it is true that:

$$Mode < Median < Mean\ (\bar{x})$$

If the skewness variable is negative, the distribution is negatively skewed and the number of Tweets is clustered at the upper end (e.g., the number of Tweets per minute is more frequent towards the higher scores) (Field, 2005). When this is the case, it is true that:

$$Mode > Median > Mean\ (\bar{x})$$

The intervals (bins) for the frequency distribution follow Scott's normal reference rule (Scott, 1979):

$$h = \frac{3.5sd}{n^{\frac{1}{3}}}$$

h … length of a bin
sd … standard deviation
n … sample size (number of Tweets)

The kurtosis shows how flat or pointy the distribution is. We distinguished between leptokurtic distributions (e.g., the distribution is relatively thin; thus, the standard deviation is small relative to the mean) and platykurtic (e.g., the distribution is relatively bright; thus, the standard deviation is more spread out relative to the mean). The further the value is from zero, the more likely it is that the data are not normally distributed (Field, 2005).

To see if the sample (number of Tweets) is representative for the population of the data (number of all Tweets which mention the EURO STOXX50 for the given time) we calculated the standard error $\sigma_{\bar{x}}$ (Field, 2005):

$$\sigma_{\bar{x}} = \frac{sd}{\sqrt{n}}$$

$\bar{x}$ … sample mean

A large standard error means that there is a lot of variability between the mean of the collected sample and the mean of the unobservable population, such that the sample we have collected might not be representative for the population. A small standard error indicates that most likely the sample mean is similar to the unobservable population mean, so it is likely that this sample is an accurate reflection of the population (Field, 2005).

Another approach to assess the accuracy of the sample mean as a good estimation of the mean of the population is to calculate confidence intervals for the true value of the mean. In other words, this indicates whether the sample size of the data set gives a fair estimation of the true unobservable mean of a population. With probability 1-α, μ (the true unobservable mean) lies between the lower and upper confidence bounds of the sample mean. Since the

standard deviation of the population is unknown, we estimated it with the standard deviation of our sample *sd* (Brannath, Futschik, & Krall, 2010):

$$\bar{x} - Q_{n-1}^{(t)}\left(1-\frac{\alpha}{2}\right)\frac{sd}{\sqrt{n}} \;\leq\; \mu \;\leq\; \bar{x} + Q_{n-1}^{(t)}\left(1-\frac{\alpha}{2}\right)\frac{sd}{\sqrt{n}}$$

Q_{n-1}…degree of freedom

If the sample mean represents the true unobservable mean, the confidence interval of the mean should be small, because with a probability of 1-α, the confidence interval contains the true mean (Field, 2005). For this research, we used the common used α of 0.05.

3.2.1.2.2 Test of normal distribution

The Kolmogorov-Smirnov test examines the sample for normal distribution. A normal distribution helps to establish proof for further statistical tests. This test compares the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the p-value is smaller than 0.05 (for a given level of α = 0.05), the test is significant. This means that the distribution of the sample is significantly different from a normal distribution, thus the distribution of the sample is non-normal. If the p-value > 0.05, the test is not significant and indicates that the sample is not different from a normal distribution, and thus it is probably normally distributed. To confirm normality or non-normality graphically, a Q-Q plot is produced. If the data are normally distributed, the observed values should fall exactly along the straight line in the chart (Field, 2005).

For further test statistics, it is necessary to transform the data when the data do not show a normal distribution. The variable “number of Tweets” was transformed and tested to see if it fits a normal distribution (Field, 2005). Since the scores of the variable “number of Tweets” are positive ($x_i > 0$) there is no problem with special functions (e.g., log function):

- Log transformation ($\log(x_i)$): The logarithm helps to reduce positive skew because it squashes the right tail of the distribution.
- Square root transformation ($x_i^{(1/2)}$): The square root transformation will bring any large scores closer to the center. Thus, this can be a useful way to reduce positively skewed data.

- Reciprocal transformation ($1/x_i$): This transformation divides one by each score and reduces the impact of large scores, thus the transformed variable will have a lower limit of zero.

3.2.2 Correlation of number of Tweets and index indicators

Hypothesis two investigates the correlation between the number of Tweets and the index indicators. We sought to uncover a correlation (a linear relationship) between the variable "number of Tweets" and the variable "EURO STOXX 50 prices". These two variables could be positively related, which would mean that a high number of Tweets goes together with a high EURO STOXX 50 index price. They could be negatively related, which would mean that a high number of Tweets shows a rather low EURO STOXX 50 price index. And finally, they could be not related at all (Field, 2005).

We sought to investigate the correlation on three different fields:

(1) Correlation of the number of Tweets per minute and the correspondent EURO STOXX 50 index price
(2) Correlation of the number of Tweets per minute and the absolute price change from one minute to another of the EURO STOXX 50 index price
(3) Correlation of the number of Tweets per minute and the relative price change from one minute to another of the EURO STOXX 50 index price

The first allows us to gauge the market movement of the EURO STOCK price and compare it with Twitter movements, so as to determine whether or not there is a correlation between Tweet volume and index price. The second two fields, on the other hand, indicate change with respect to fluctuations of the index and its relationship with Twitter activity. A large delta of change underlies a higher volatility of the security, which can lead to a dramatic change in the price over a short time in both directions. A lower spread of the changes underlies a lower volatility of the index and the security value may not fluctuate dramatically (Investopedia, 2017).

Besides the statistical testing of the correlation coefficients, we looked at the different scatterplots to get information about the general trend of the data. The scatterplots show

the number of Tweets as the independent variable (x-axis) and the EURO STOXX index prices as the dependent variable (y-axis).

3.2.2.1 Pearson's correlation coefficient

The Pearson product-moment correlation coefficient *(r)* is defined by:

$$r = \frac{cov_{xy}}{sd_x sd_y} = \frac{\Sigma(x_i - \overline{x})\overline{(y_\iota - \overline{y})}}{(n-1)sd_x sd_y}$$

This equation ends up with a value between -1 and +1. A coefficient *r* of -1 indicates a perfect negative relationship. Thus, if one variable is high, the other variable is low. Conversely, a coefficient of +1 indicates a perfect positive relationship and says that as one variable is high the other is high as well. A coefficient of 0 indicates no linear relationship. The correlation coefficient is also a standardized measure for the size of an effect. Values of +/-.1 represent a small effect, +/-.3 is a medium effect and +/-.5 is a large effect (Field, 2005).

Pearson's correlation requires that the data are measured at an interval or ratio level. For the test statistics to be valid as to whether the correlation coefficient is significant, data must be normally distributed. Although the correlation coefficient cannot give any information about causality, we take the squared correlation coefficient (R^2) to describe the amount of variability in one variable that is explained by the other (Field, 2005).

3.2.2.2 Daily correlations and time series analysis

When we have statistical significant correlations of the variables, it helps to plot the variables in a time series to discuss the movements and shifts of the variables and their interaction with each other. We will plot the outlier days and discuss time series over a longer period (e.g., September 13-20, 2016). On this basis, we will consider daily (intraday data) correlations of the EURO STOXX and the number of Tweets to see if there is a significant correlation of data points within a day.

3.2.2.3 Spearman's correlation coefficient

We used Spearman's correlation coefficient, a non-parametric correlation coefficient, as we do not have normally distributed data. For this test, we ranked the data and then applied

Pearson's equation. When using the Spearman's correlation coefficient (r_S) we rank the data by change from the previous score:

Rank	Range $rx_S = [(x_i - x_{i-1})/n]*100$
1	[max r_S; 0.3]
2	[0.29;0.2]
3	[0.19;0.1]
4	[0.09; min r_S]

Table 2: Rank order for Spearman's correlation coefficient

3.2.2.4 Partial correlation

In our data set the constant variable is time. Since we investigated the correlation between the variables "number of Tweets per minute" and "EURO STOXX" (including its changes from one minute to another) we were interested to see how these variables interact beyond time. Since we investigated the correlation between the variables "number of Tweets" and "EURO STOXX" and the variable "time," we needed to consider the influence of the time on these two variables. Conducting a partial correlation helps to get an understanding of the overlaps to find out the size of the unique portion of variance. The third variable is a controlling variable of the correlation effect. The partial correlations can be done for dichotomous variables (Field, 2005).

3.2.3 Forecasting model

Hypothesis three concerns the concept of a forecasting model. We selected the highest correlation pair (either the correlation between the variables "number of Tweets" and the "EURO STOXX 50 price per minute" or the correlation between "number of Tweets" and the relative or absolute price change from one minute to another). We used a linear regression analysis to find a predictive model that fit the data. This model was used to predict future values for the dependent variable ("EURO STOXX 50 price per minute" or relative/absolute price change). This means that we sought to find a model that predicts the EURO STOXX 50 prices on a one-minute interval basis (the dependent variable) or predicts the absolute or the relative price change of the EURO STOXX 50 index from one minute to another, underlying the independent variable "number of Tweets per minute." We built our model

based on the intraday (minute-to-minute) input data from September 13 to October 14, 2016. To test our model, we used the observed data from October 19 and compared them with the predictive data.

3.2.3.1 Method of least squares

The method of least squares is a way to find the line that best fits the data (Field, 2005). The regression model is as follows:

$$Y_i = \alpha + \beta X_i + \varepsilon_t$$

Y_t represents the dependent variable and X_{t-1} represents the independent variable. The parameters α and β stand for the regression parameters and ε indicates the error term.

The model established in this research is based on a simplified regression model based on the model used in Mao et al. (2012), who build an exogenous predicting model for a financial index using Twitter data. Mao et al.'s regression model is based on the index indicator of t-1 and the number of Tweets in t-1:

Model 1:

$$Y_t = \alpha + \beta Y_{t-1} + \gamma X_{t-1} + \varepsilon_t$$

In our model Y_{t-1} denotes the EURO STOXX 50 index price or the absolute and relative price change at time t-1 (t-1 = price a minute ago), X_{t-1} denotes the number of related Tweets at time t-1 (t-1 = number of Tweets a minute ago), α, β_i and γ_i are the regression parameters and ε indicates the error term. The assumption for this model contradicts the random walk theory, which, according to Paul Samuelson (1965), says that price changes must be unforecastable if they fully incorporate the expectations and information of all market participants. This means that the more efficient the market, the more random is the sequence of price changes. This model states that the best prediction for Y_t is the price of Y in t-1.

We compared the multiple regression model with a simple regression model of only Y_{t-1} (EURO STOXX 50 price a minute ago) and a simple regression model with only X_{t-1}:

Model 2:

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$$

Y_{t-1} EURO STOXX 50 price

α, β regression parameters

ε error term

Model 3:

$$Y_t = \alpha + \beta X_{t-1} + \varepsilon_t$$

X_{t-1} number of Tweets per minute

α, β regression parameters

ε error term

3.2.3.2 Testing of regression models

For the testing of the models, we computed predictive prices of the EURO STOXX 50 on a one-minute interval basis. Following Mao et al. (2012) we distinguished between a price increase and a price decrease. If the price increased, we denoted the value +1. If the price decreased, we denoted the value -1. If there was no price change, we denoted the value 0. We compared the predicted values with the observed values and computed the prediction accuracy and plot the comparison on a bar chart.

4 Results

This chapter describes the results of the hypotheses asked in Chapter 2, "Field of investigation and hypotheses." Each paragraph corresponds with the respective paragraph and hypothesis in Chapter 2. All tables of statistics (table of descriptive statistics, frequency table, etc.) can be found in the Appendix, in the section "Statistics."

4.1 Data mining of European Twitter data

4.1.1 Frequency of Tweets per day and text analysis

Over the observational period from September 13, 2016 at 09:00 am to October 19, 2016 at 05:30 pm (18 days) in total 373,250 Tweets have been collected (see Appendix for frequency table). Figure 5 shows the distribution of captured Tweets on a time series plot on a day-to-day basis. In regards to the distribution during the day, it is obvious that the number of Tweets in the morning (09:00 am to 12:00 pm) is for all days lower than the number of Tweets in the afternoon (12:01 to 05:31 pm). Hence, Twitter users are more active within the second half of the day. On average, we collected 20,736 Tweets per day (in average 5,274 Tweets in the morning and 15,462 Tweets in the afternoon).

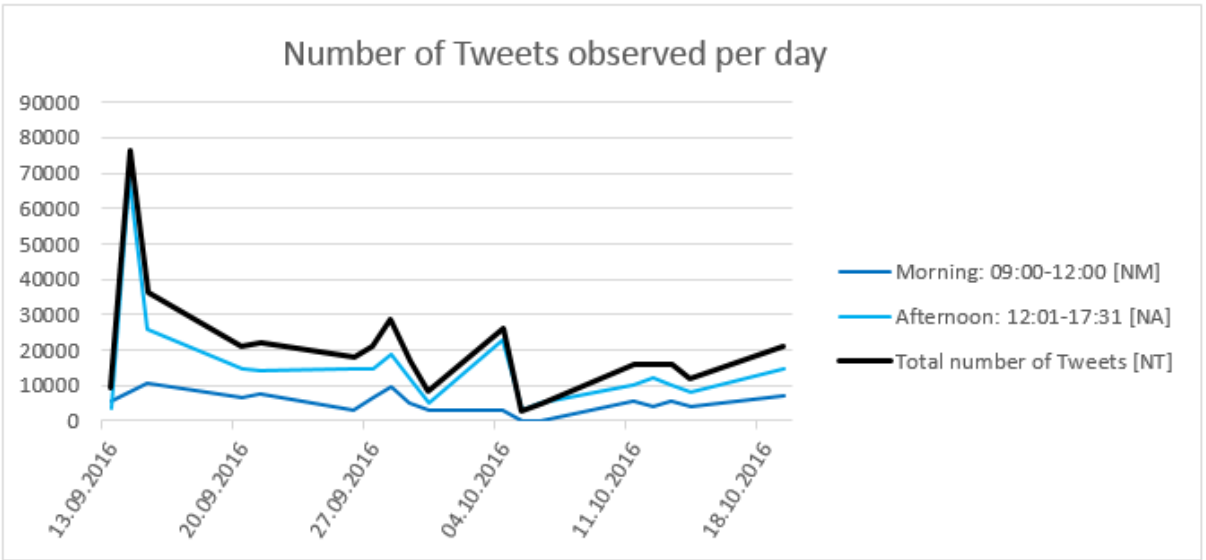


Figure 5 Time series plot

Table 16 shows the frequency of the number of Tweets observed per day and indicates the outliers. In the data set, the most significant outlier values ($Outlier_2$), indicating that a high

number of Tweets have been spread, are captured from September 14, 2016 in the afternoon to September 15, 2016 in the morning and also in the morning of September 28, 2016. This can be seen in Table 3 (see also the frequency table in the Appendix, Table 18).

Date	Total Number of Tweets [N_T]	Number of Tweets between 09:00-12:00 [N_M]	Number of Tweets between 12:01-17:31 [N_A]
14.09.2016	**76536****	8065*	**68471****
15.09.2016	36356*	**10580****	25776
28.09.2016	28559	**9598****	18961

Table 3 Outliers

* Value is higher than 2*sd (=$Outlier_1$)
** Value is higher than 3*sd (=$Outlier_2$)

When we plot the outliers of September 14 and 15, 2016 within a minute-to-minute interval time series chart, it is obvious that these data are highly variant (see Figure 6).

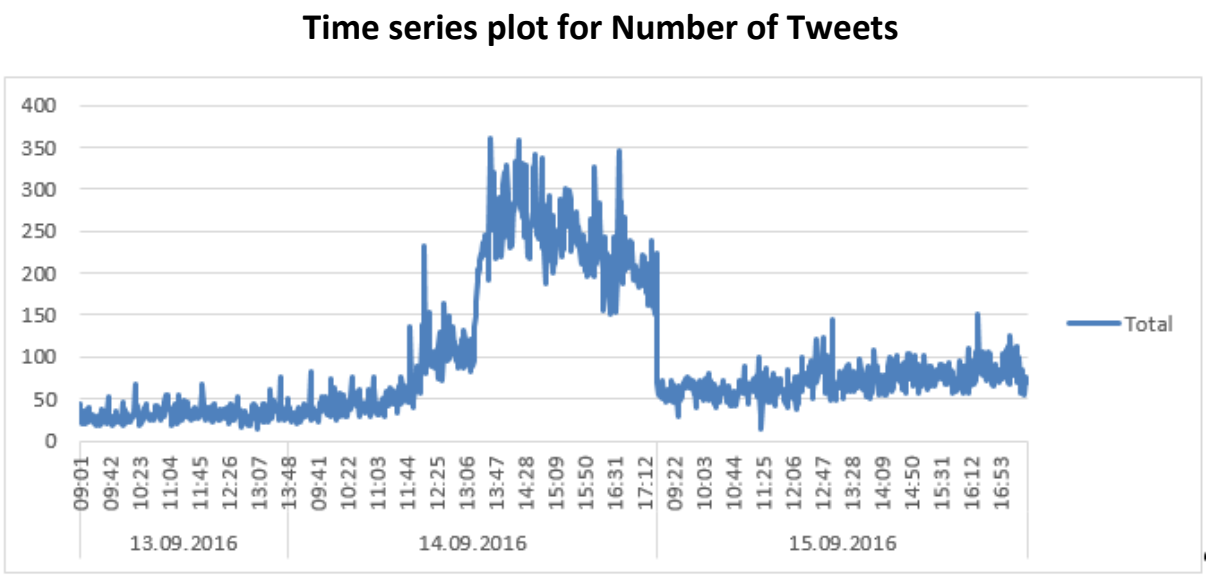


Figure 6 Time series plot between September 13-15, 2016

The text analysis of these days gives an indication of the content of the Tweets. The tables below summarize the top five most frequently used words in the tweeted messages over the days of September 14, 15 and 28 of 2016:

Most frequently used words of September 14, 2016:

#	Text	Frequency
1	Bayer	113679
2	Carrefour	6735

3	Allianz	4564
4	Repsol	3507
5	Bbva	3081

Table 4 Most frequently used words of September 14, 2016

Most frequently used words of September 15, 2016:

#	Text	Frequency
1	Bayer	72373
2	carrefour	13544
3	allianz	9885
4	Bbva	9241
5	Vinci	6837

Table 5 Most frequently used words of September 15, 2016

Most frequently used words of September 28, 2016:

#	Text	Frequency
1	Bbva	39972
2	deutschebank	18573
3	bayer	8581
4	Axa	7416
5	Carrefour	6796

Table 6 Most frequently used words of September 28, 2016

When considering these data, we can easily observe that the content of the Tweets contains most frequently the words, among others, “Bayer” (for Bayer AG[2]), “BBV” (for Banco Bilbao Vizcaya Argentaria S.A.[3]), “Deutsche Bank” (for Deutsche Bank AG[4]) or “Carrefour” (for Carrefour S.A.[5]). The mentions of these keywords led to the increase in the number of Tweets, hence to the outliers in the data.

[2] see https://www.bayer.com/
[3] see https://www.bbva.es/particulares/index.jsp
[4] see https://www.deutsche-bank.de
[5] see http://www.carrefour.com/

4.1.2 Variable number of Tweets per minute

The data represent a big data set, for which we explored the collected number of Tweets per minute. Figure 7 shows an extract of the movement of the variable “number of Tweets per minute” from September 20–22 and September 26–27, 2016. The x-axis indicates the time; the y-axis indicates the number of Tweets per minute. The charts allow us to identify the outliers (up to the value 160 within the captured time range) easily. Table 18 shows the frequency table of the variable. The minimum value is six, the maximum value is 380. Thus, the maximum number of Tweets published within a minute that contained the related EURO STOXX 50 keywords and/or company names was 380.

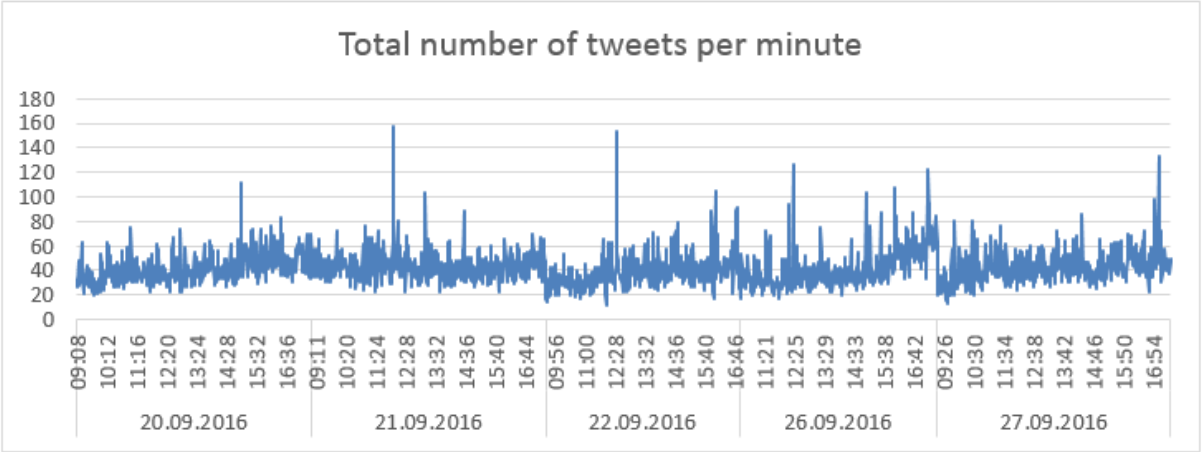


Figure 7 Extract of the Tweets movement over time

The box-and-whisker plot (see Figure 8) shows the median of the variable “number of Tweets per minute” was 42. The first quartile began at value 33, and the third quartile at value 56. Values below 74 Tweets per minute are indicated as outliers.

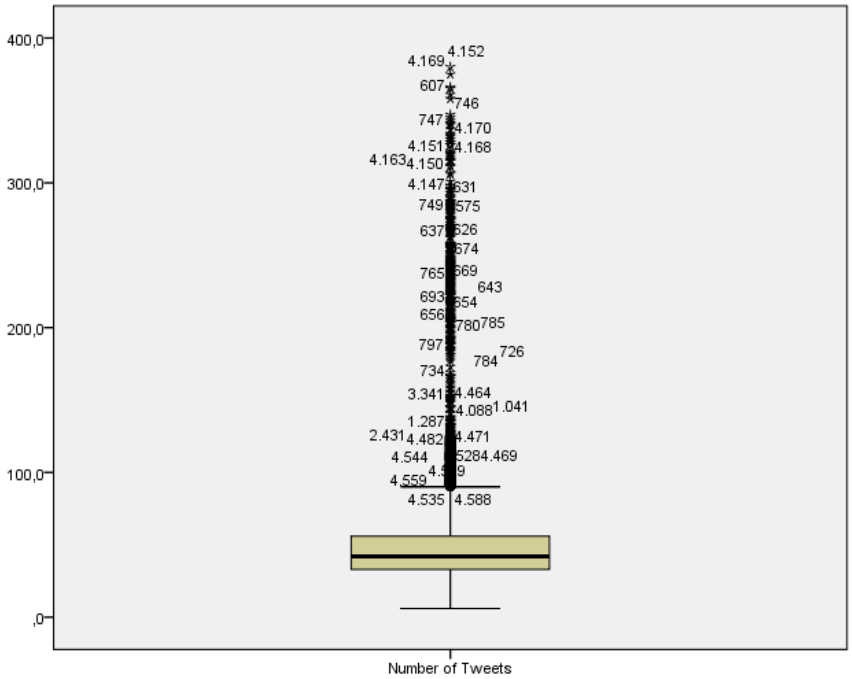


Figure 8 Number of Tweets per minute–box-and-whisker plot

4.1.2.1 Descriptive statistics: univariate analysis

The histogram of the variable "number of Tweets per minute" (see Figure 9) shows that the variable is positively skewed, with a skewness of 3.711. This means that that the data are more clustered between zero and the lower end of the mean. In fact, the range of the most frequent observations lies in the bin of 31-40 with 1,892 observations (see Appendix for frequency, Table 18) and a mode of 35. In other words, in our data set we see 1,892 observations for 31-40 Tweets in a minute.

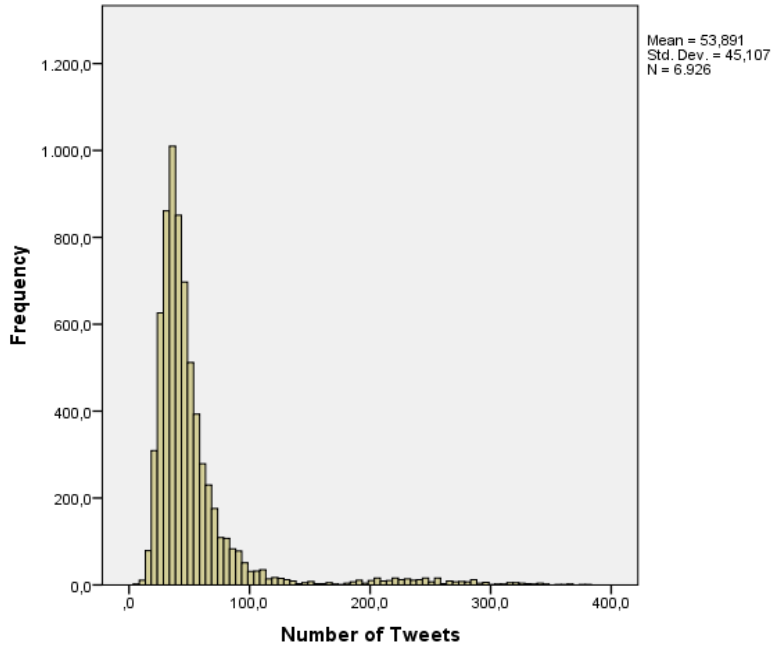


Figure 9 Histogram of the variable "number of Tweets"

The histogram shows a relatively thin distribution; hence the data include high peaks, with a kurtosis value of 15.607. The mean of the data lies at 53.89 with a standard deviation of 45.11. When we have a small standard deviation relative to the mean, we speak of a leptokurtic distribution. The high value of the kurtosis also indicates that the distribution is more likely to be not normally distributed.

The confidence interval of the mean has a probability of 95% between 52.829 and 54.954. The standard error of the mean is, with 0.542, relatively high. The sample of our data might not be representative for the population, i.e., the mean of our sample may not represent the mean of the population even though the confidence interval is relatively small.

4.1.2.2 Test of normal distribution

The test statistics of the Kolmogorov-Smirnov test show that the variable “number of Tweets” is not normally distributed. The p-value is below 0.05, which means the sample is significantly different from a normal distribution (see Figure 10).

The Q-Q plot of the observed values shows that they do not fall along a straight line. Thus, Figure 11 shows that the sample size for neither the variable “number of Tweets” nor its transformed variables is normally distributed.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Number of Tweets is normal with mean 53,9 and standard deviation 45,107.	One-Sample Kolmogorov-Smirnov Test	,000[1]	Reject the null hypothesis.
2	The distribution of Log is normal with mean 1,66 and standard deviation 0,228.	One-Sample Kolmogorov-Smirnov Test	,000[1]	Reject the null hypothesis.
3	The distribution of Squareroot is normal with mean 6,99 and standard deviation 2,249.	One-Sample Kolmogorov-Smirnov Test	,000[1]	Reject the null hypothesis.
4	The distribution of divided is norma with mean 0,02 and standard deviation 0,011.	One-Sample Kolmogorov-Smirnov Test	,000[1]	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1] Lilliefors Corrected

Figure 10 Kolmogorov-Smirnov for transformed variable "Number of Tweets"

However, for further analysis we use the log(10) transformation, since this transformed variable most likely describes a normal distribution (see discussion of Q-Q plots in Figure 11).

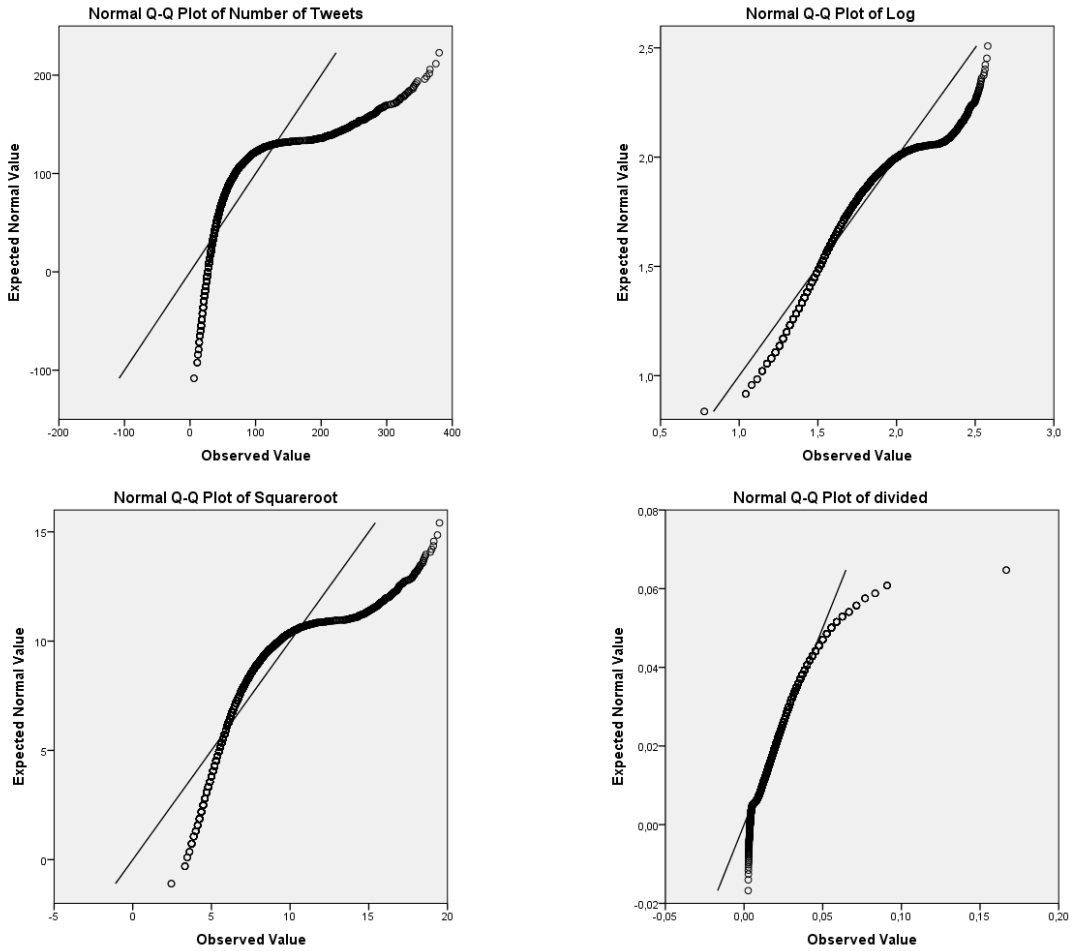


Figure 11 Q-Q plots for the transformed variable "number of Tweets"

4.2 Correlation of number of Tweets and EURO STOXX50

4.2.1 Pearson's correlation coefficient

Table 7 shows the correlation table of both the variable "number of Tweets per minute" and its linear relationship to the EURO STOXX 50 index price at the same time (intraday one-minute sequence). The columns "absolute change" and "relative change" indicate the absolute and respective relative price change of the EURO STOXX 50 from one minute to another.

The correlation of the variable "number of Tweets" and the corresponding EURO STOXX 50 index price is negative, with a Pearson correlation coefficient of -0.241. This means that when the variable "number of Tweets" is high, the variable "EURO STOXX 50" is rather low.

		EURO STOXX 50	Δ (abs change)	Δ% (rel change)
Number of Tweets	Pearson Correlation	-.241**	.043**	.051**
	Sig. (2-tailed)	.000	.000	.000
	Sum of Squares and Cross-products	-2103427.995	11422.306	4.261
	Covariance	-303.744	1.650	.001
	N	6926	6925	6923

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7 Correlation of number of Tweets per minute and EURO STOXX index per minute

The scatterplot of the variable "number of Tweets per minute" on the x-axis versus the variable "EURO STOXX 50" on the y-axis (see Figure 12) reflects the correlation of the market movements and compares it with the Twitter movements. The data points representing a number of Tweets per minute under 100 are distributed between a wider range of the EURO STOXX 50 price scale. However, when the volume of the number of Tweets is high (x-axis) the data points obviously lie on the lower end of the EURO STOXX 50 price range.

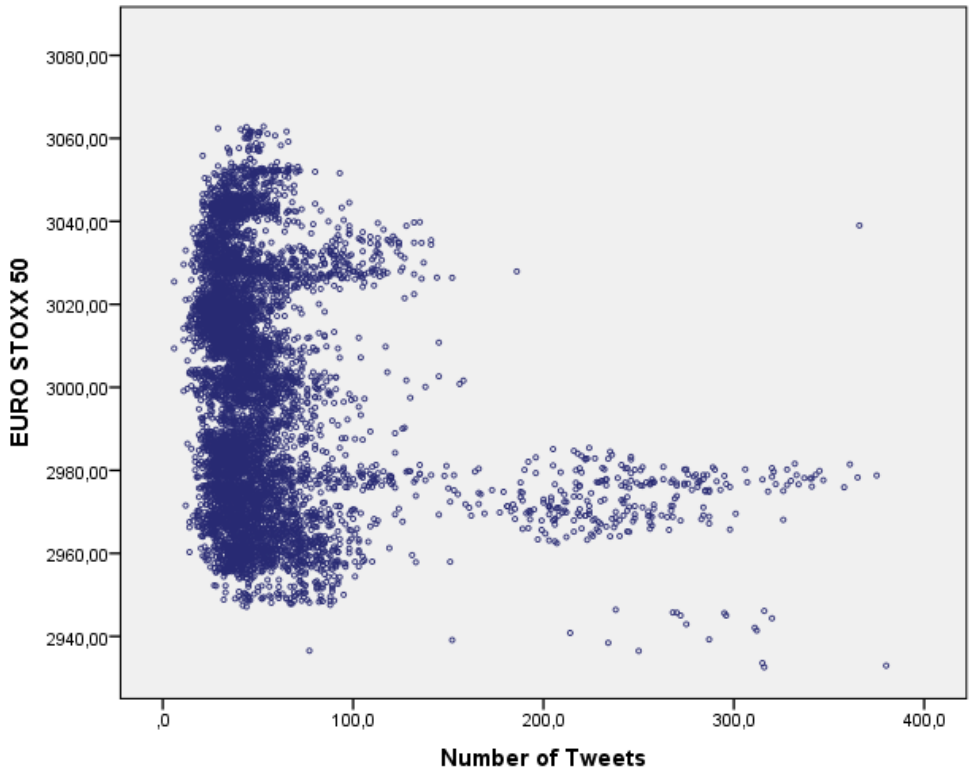


Figure 12 Scatterplot EURO STOXX 50 with number of Tweets per minute

The following graphs treat the question of whether we can see a correlation of the number of Tweets and the price change of the EURO STOXX 50—that is, whether a high number of Tweets significantly correlates with a high price difference on both, the absolute price change per minute (see Figure 13) and the relative price change per minute (see Figure 14). Since the correlation coefficient of the number of Tweets per minute and the absolute price change is 0.043, and 0.051 for the relative price change, the data do not show a correlation between these two variables. Neither of the graphs give this indication.

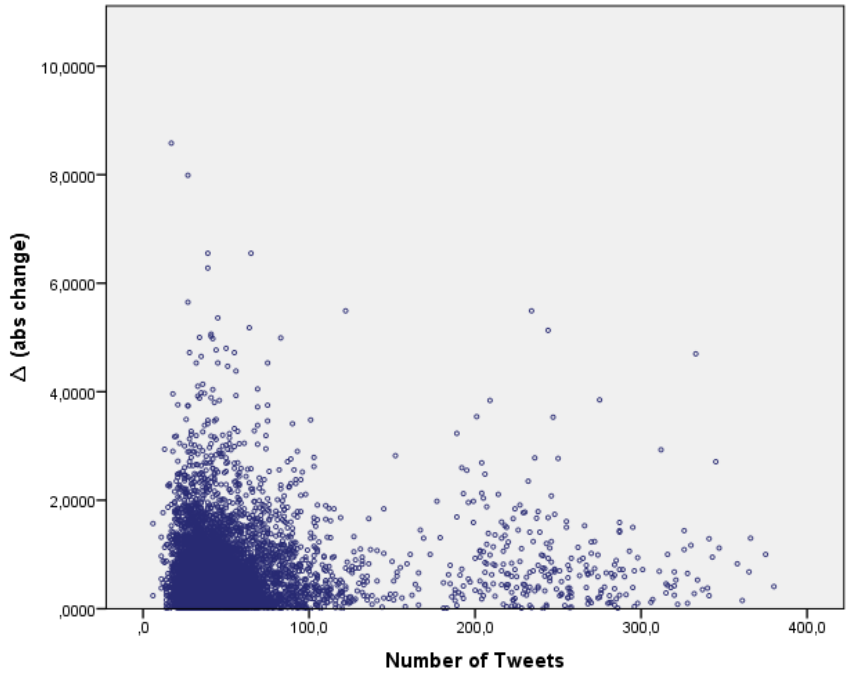


Figure 13 Scatterplot of the absolute change of EURO STOXX from minute to minute and the variable "number of Tweets per minute"

The graphs (Figure 13 and Figure 14) show that the price change of the EURO STOXX does not fluctuate greatly. Even with a high number of Tweets the data points fall on the lower price change end (up to 0.1% from one minute to another).

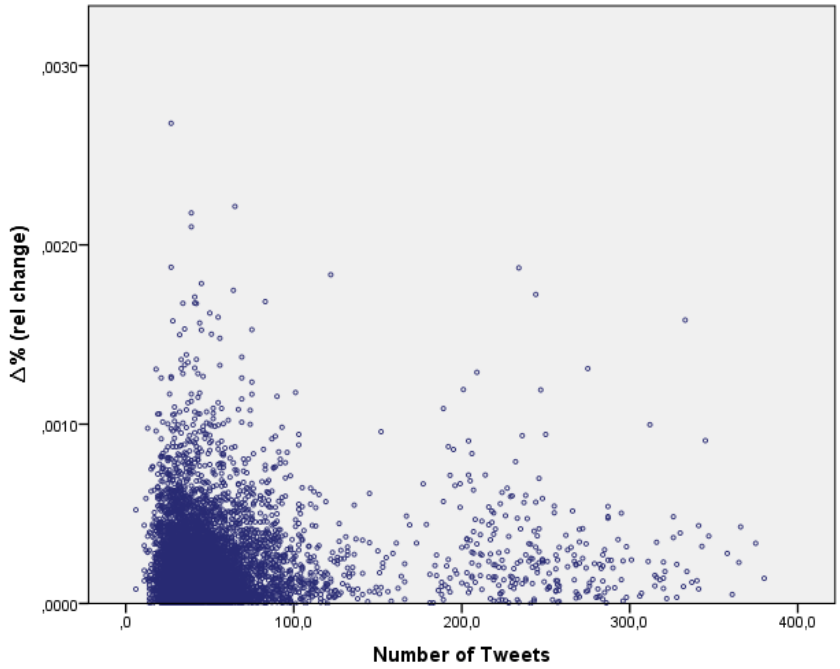


Figure 14 Scatterplot of relative change of EURO STOXX per minute and number of Tweets per minute

4.2.2 Daily correlations and time series analysis

Between September 13 and 15, 2016, we observed a peak in the number of Tweets. From the frequency table in Chapter 3.2.1.1, "Frequency of Tweets per day and text analysis," we know that at the number of Tweets was very high on September 14, 15, and 28, 2016. When we compare these days with the EURO STOXX frequency table (see Table 8) we see that the EURO STOXX 50 price on these days was very low compared to the other days.

Date	Min of EURO STOXX 50	Max of EURO STOXX 50	Average of EURO STOXX 50	*StdDev of EURO STOXX 50*	*Var of EURO STOXX 50*
13.09.2016	3007.0	3032.3	3015.3	*4.982*	*24.825*
14.09.2016	**2962.4**	2989.1	2975.6	*5.193*	*26.970*
15.09.2016	**2947.6**	2977.8	2962.8	*6.242*	*38.967*
20.09.2016	2959.4	2984.6	2975.2	*4.669*	*21.801*
21.09.2016	2980.8	3009.9	2996.7	*7.176*	*51.502*
26.09.2016	2974.3	2990.0	2982.7	*3.241*	*10.506*
27.09.2016	2947.1	2996.3	2965.0	*11.999*	*143.965*
28.09.2016	**2984.0**	3010.9	2998.4	*5.840*	*34.103*
29.09.2016	2996.2	3031.9	3017.0	*7.660*	*58.683*
30.09.2016	2932.6	2978.9	2952.4	*17.756*	*315.282*
04.10.2016	3013.9	3040.3	3024.7	*5.799*	*33.629*
05.10.2016	3026.0	3031.0	3028.2	*1.289*	*1.663*
06.10.2016	3025.6	3034.0	3028.9	*2.058*	*4.236*
11.10.2016	3019.6	3050.8	3037.6	*6.477*	*41.950*
12.10.2016	3005.5	3023.2	3014.9	*4.543*	*20.639*
13.10.2016	2954.4	2986.0	2967.5	*5.322*	*28.319*
14.10.2016	2997.4	3035.0	3020.8	*11.298*	*127.644*
19.10.2016	3038.6	3062.8	3046.7	*5.624*	*31.626*
Grand Total	**2932.6**	**3062.8**	**2999.0**		

Table 8 EURO STOXX 50 frequency table from September 13 to October 19, 2016

When we plot the variable "number of Tweets" and the variable "EURO STOXX 50" on a time plot, it is obvious that the EURO STOXX 50 collapsed from September 13 to 14 and stayed on average at a level of about 2.970. On the other side the number of Tweets had a peak on September 14 in the afternoon and continued to be very high on September 15 in the morning.

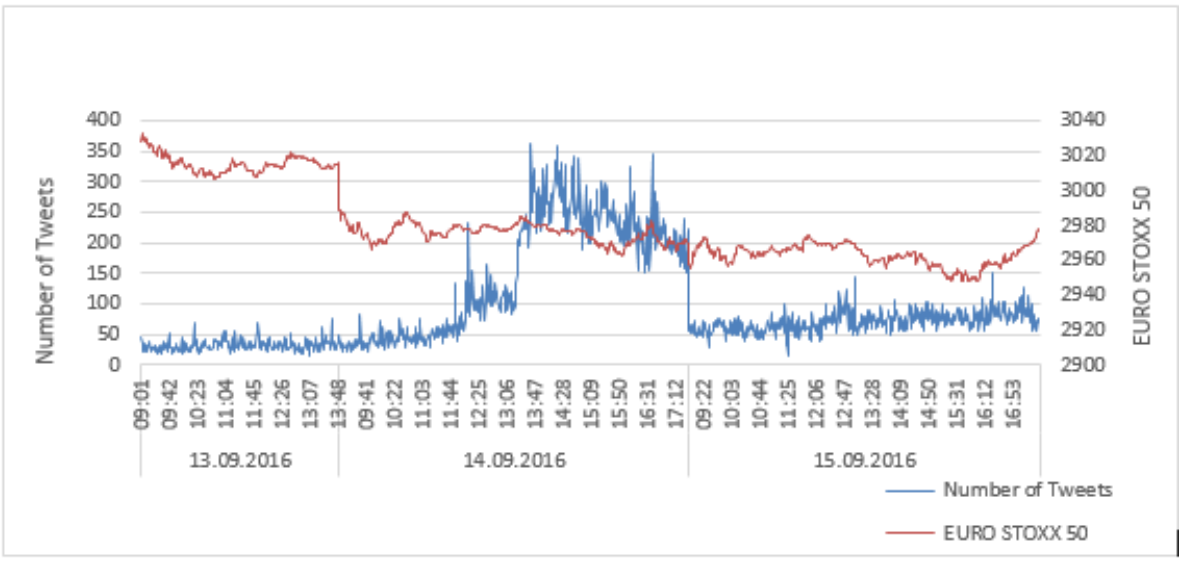


Figure 15 Time series analysis extract from September 13-15, 2016

Taking into consideration the daily correlation (see Table 9) of the variables "number of Tweets" and "EURO STOXX 50" we see that some days are more highly correlated than others. The following time series plots discuss these days, on which there was a significantly high correlation, both positive and negative.

Date	Pearson's Correlation	N
13.09.2016	**-.236****	288
14.09.2016	**-.272****	511
15.09.2016	**-.189****	511
20.09.2016	.106*	502
21.09.2016	0.051	503
26.09.2016	**-.236****	424
27.09.2016	**-.301****	500
28.09.2016	**-.130****	511
29.09.2016	**-.265****	393
04.10.2016	**.775****	418
05.10.2016	-0.061	48
06.10.2016	-0.100	91
11.10.2016	.120*	449
12.10.2016	**-.505****	477
13.10.2016	**-.175****	428
14.10.2016	**.413****	355

19.10.2016		**.290****	488

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 9 Daily correlation of number of Tweets and EURO STOXX 50

For October 4, 2016, the number of Tweets and the EURO STOXX 50 are positively correlated with a coefficient of 0.775 and a significance level of 0.01. In the time series plot we can see that the movement of the two variables is nearly parallel. At around 16:52 there was a peak for both the number of Tweets and the EURO STOXX 50 (see Table 10).

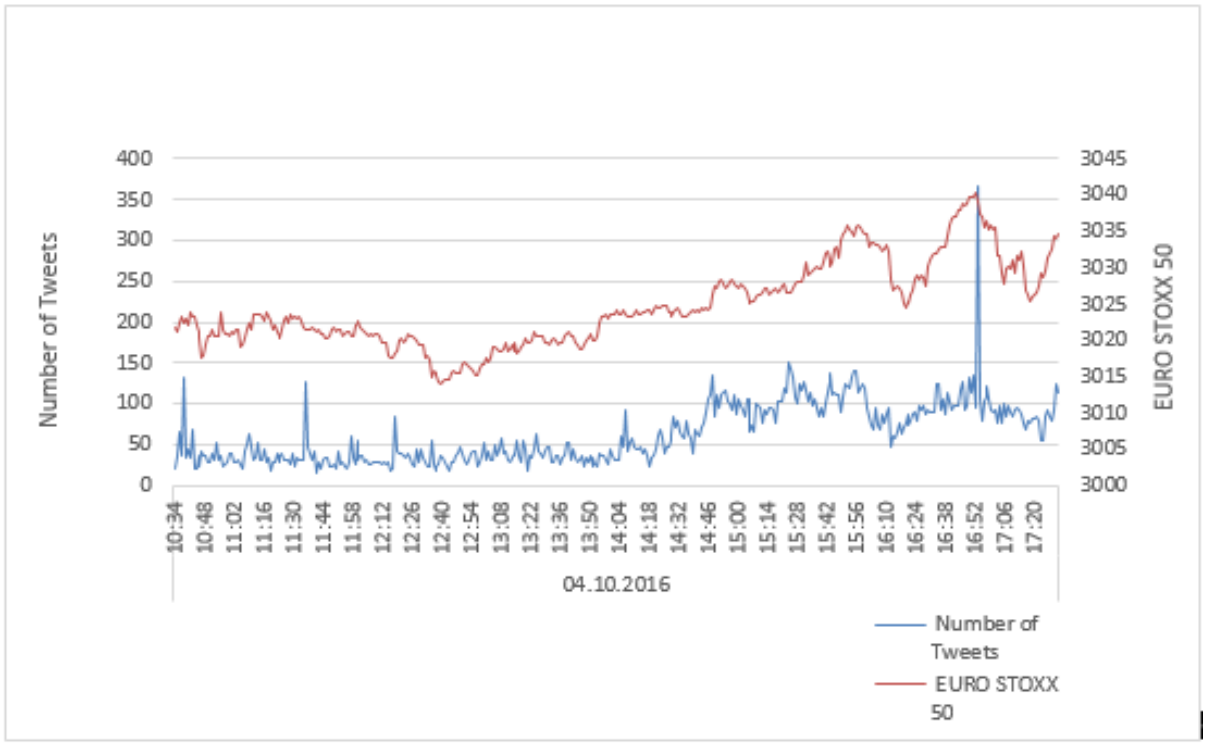


Table 10 Time series plot of October 4, 2016

Table 11, on the other hand, shows the opposite. The correlation coefficient of this day is -0.505, which means that when the number of Tweets was high, the EURO STOXX price was rather low.

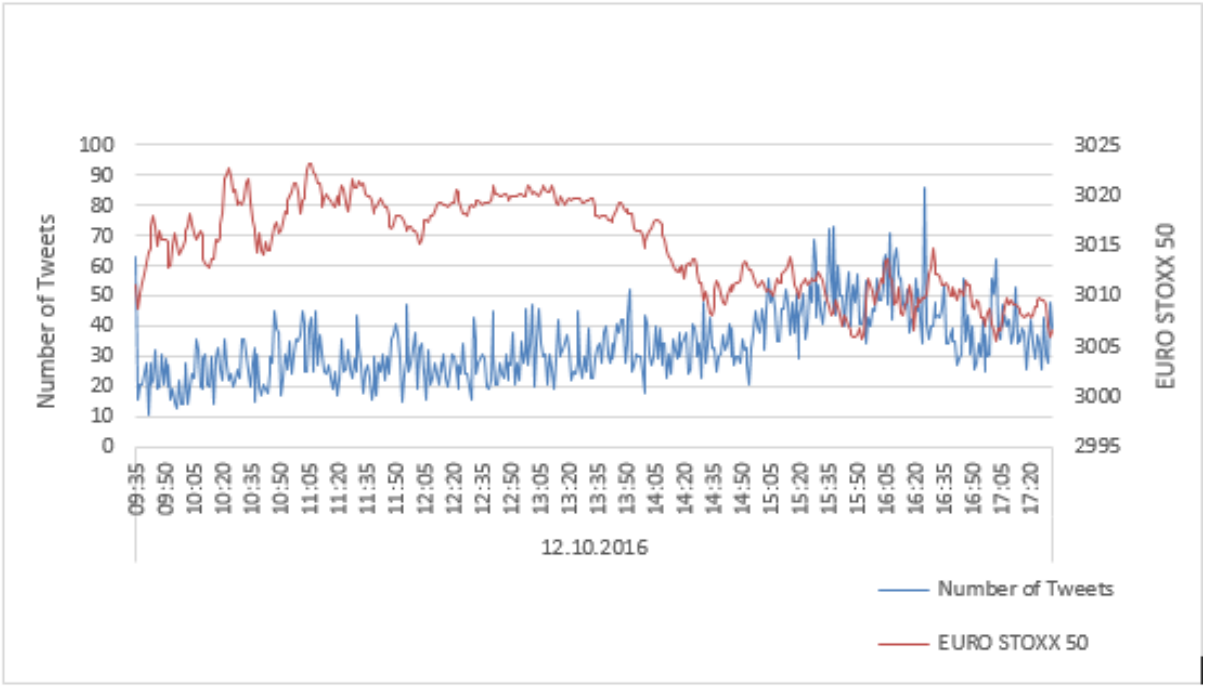


Table 11 Time series plot of October 12, 2016

4.2.3 Spearman's correlation coefficient

The following discussion further investigates different ways of calculating correlation coefficients. Since our data were not normally distributed, it helps to apply the Spearman's correlation methodology.

Spearman's Correlations

			Number of Tweets	Δ (abs change)	Δ% (rel change)	EURO STOXX 50
Spearman's rho	Number of Tweets	Correlation Coefficient	1,000	.016	.019	-.247**
		Sig. (2-tailed)	.	.188	.117	.000
		N	6926	6925	6923	6926

**. Correlation is significant at the 0.01 level (2-tailed).

Table 12 Spearman's correlation coefficient

The Spearman correlation coefficient for the variable "number of Tweets per minute" and the variable "EURO STOXX 50" is, with -0.247, slightly higher than the Pearson correlation coefficient. Also, the correlation coefficient of the absolute and relative change between the

EURO STOXX 50 and number of Tweets is, according to the Spearman's correlation coefficient, not correlated, with the values 0.016 and 0.019.

4.2.4 Partial correlation

When computing partial correlation, taking into consideration that the variable "time" is the controlling variable, we receive a correlation coefficient of -0.061. This means that the variables are also correlated with the time and bias the result. The correlation coefficient is, with -0.061, close to 0, which means that excluding the variable "time" leads to almost a non-correlation of the variables "number of Tweets" and "EURO STOXX 50."

Partial correlations

Control Variables			EURO STOXX 50	Number of Tweets
Date	EURO STOXX 50	Correlation	1,000	-,061
		Significance (2-tailed)	.	,000
		Df	0	6923

Figure 16 Partial correlation with time as controlling variable

4.3 Forecasting model

Since the variable "number of Tweets" and the variable "EURO STOXX 50" have the highest correlation, we ran the regression analysis with this variable pair within the three regression models:

Dependent Variable: EURO STOXX 50

	Model 1	Model 2	Model 3
Constant	4.611	6.540	3002.548
	(2.556)	(2.476)***	(0.476)***
EURO STOXX 50 $_{t-1}$	0.998	0.998	
	(0.001)***	(0.001)***	
Number of Tweets $_{t-1}$	0.001		-0.131
	(0.000)***		(0.007)***
Observations:	6436	6436	6436
R-squared:	0.996	0.996	0.057
F-statistic:	729794.381	1457768.668	390.879
Standard error of the estimate	1.69	1.69	24.78

* significant at 10% level; ** significant at 5% level; *** significant at 1% level
(Standard errors in parentheses)

Table 13 Output of regression analysis

The percentage of the response variable (R-squared) that is explained by Model 3 is very low. The model explains 5.7% of the variability of the response data around their mean.

Model 1 and 2 show a R-squared of more than 99%. All parameters are significant at a 1% significance level.

Testing of the regression models

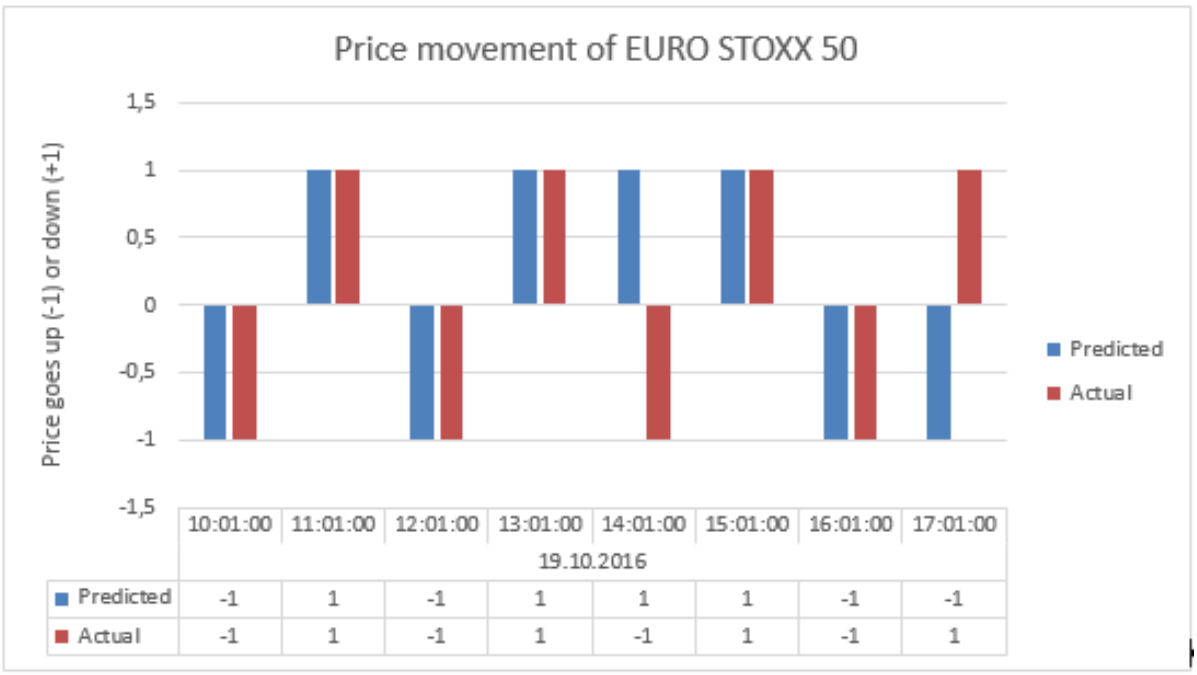
We tested the three models at October 19, 2016 within a one-minute time interval of the EURO STOXX 50 price change. We predicted the EURO STOXX 50 with the three models and computed the cases in which the predicted price movements fit to the observed price movements (see Table 14).

	Model 1	Model 2	Model 3
Number of cases in which predicted price movement is identical to observed price movement	261	259	245
Number of cases in which predicted price movement differs from observed price movement	225	227	241
Total number of cases:	486	486	486
Accuracy of prediction:	53.70%	53.29%	50.41%

Table 14 Predicted and actual change of EURO STOXX 50 (one-minute time interval)

The result shows that Model 1 (this model includes the number of Tweets) gives a slightly higher prediction accuracy than Model 2 (0.41 percentage points higher than the prediction accuracy of Model 2), considering that the standard error of Model 1 is 1.69. Only having the variable “number of Tweets” in focus (Model 3) gives an accuracy level of 50.41%.

Figure 17 shows the price movement of the EURO STOXX 50 over a one hour interval. Actual price movements are marked in red. The graph reflects the prediction values from Model 3 (predicted values are marked in blue). If the price goes down from one minute to another it is indicated with -1. If the price goes up, it is indicated with +1.



	10:01:00	11:01:00	12:01:00	13:01:00	14:01:00	15:01:00	16:01:00	17:01:00
	19.10.2016							
Predicted	-1	1	-1	1	1	1	-1	-1
Actual	-1	1	-1	1	-1	1	-1	1

Figure 17 Excerpt of price movement of EURO STOXX 50, observed and predicted values

5 Conclusion

5.1 Data mining of European Twitter data

With the given keywords observed and collected within the period from September 13 to October 19, 2016, it was possible to collect on average 20,735 Tweets per day, which is an impressive set of data. Our data set shows that Twitter users tend to Tweet in the afternoon rather than in the morning, even though the content is related to the economic and financial information of companies, which spreads throughout the entire day. There are obviously many private users who are more active in the afternoon. Most of the information has its source in company accounts and public relation departments, political accounts or institutions of the stock markets. This fact leads us to the assumption that many private Twitter users comment and publicize their opinions on economic information. Thus, we see a significant interest in spreading information about the economy, as such, and the financial markets from private users on European Twitter accounts.

5.1.1 The Bayer and Monsanto merger: September 14, 2016

The assumption that private Twitter users spread relevant economic information is also evident in the section “Text analysis of the most frequent words.” On September 14, 2016, Bayer and Monsanto announced their merger agreement under which Bayer will acquire Monsanto for USD 128 per share in an all-cash transaction (Bayer AG, 2017). The keyword “Bayer” was mentioned more than 113,000 times over the course of September 14, 2016. Figure 6 shows this movement on a time series plot. The number of Tweets was significantly higher due to the increased mentions of Bayer-related content. When considering the content of the Tweets, we can see that the reaction to this merger was mostly negative. Below there is an excerpt of Tweets posted during the day (see Figure 18).

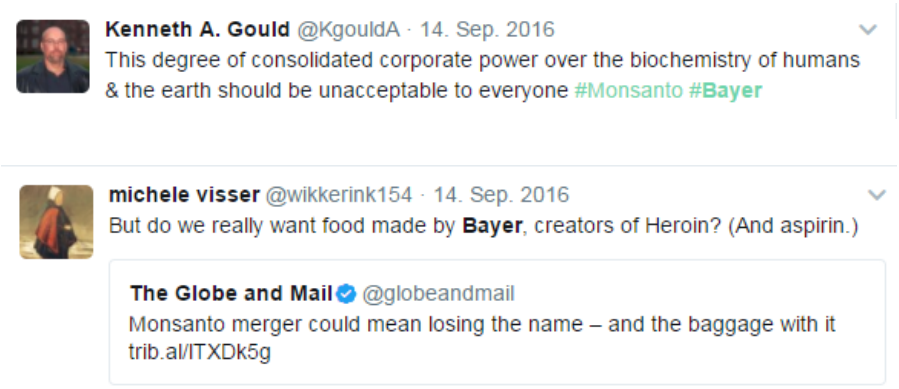


Figure 18 Excerpt of Tweets mentioning Bayer on September 14, 2016

5.1.2 Deutsche Bank and its increase in Tweets on September 28, 2016

Another comparison of the mood on social media is reflected on September 28, 2016. On that day, there were more than 18,500 mentions of the keyword "Deutsche Bank" on Twitter. The media published that the German government was providing a rescue plan for Deutsche Bank after the threat of a $14 billion lawsuit from the US Department of Justice (Die Zeit/Wirtschaft, 2017). Below are some Tweets commenting on the situation.

Figure 19 Tweets about Deutsche Bank during the day of September 28, 2016

Since we can follow many sources and information on social media, and the information is spread at a very fast pace, it is natural to think that social media has an influence on the change of stocks. In this study, we have not performed any sentiment analysis to assess the mood on social media. Further research should focus on text sentiment analysis. With this method, someone could understand the market mood and predict stock markets accordingly.

5.2 Correlations of Tweets and the EURO STOXX 50

When we focus on the number of Tweets mentioning EURO STOXX 50 companies, we see a significant correlation between the number of Tweets posted within a minute and the intraday minute price of the EURO STOXX 50. The correlation coefficient has, with -0.241, a weak negative correlation. This leads to the assumption that Twitter users tend to spread information which is negatively related to the companies, due to their cautiousness. We know that investors act in a manner that is risk-averse (e.g., investors expect a higher return for a higher volatility) and focus on the security's risk before investing. The high number of Tweets spread when negative information arises strengthens the assumption that investors would act cautiously and sell their securities, which leads to a decrease in prices.

5.3 Forecasting of EURO STOXX 50 prices and outlook

The model forecast's accuracy, with the parameters price of EURO STOXX 50 of the previous minute and number of Tweets of the previous minute, is with 0.41 percentage points slightly higher than just having the parameter EURO STOXX 50 price of the previous minute. This leads to the assumption that the additional parameter "number of Tweets" improves the model only on a very small level underlying the data set. This model is heavily reliant on past

performance and has a standard error of 1.69. With this in mind, it is difficult to make a statement about the prediction accuracy compared to Model 2.

For further investigations and a more accurate forecasting model, it is recommended that researchers observe data for a longer period (e.g., observation of Tweets over a six month period).

6 Summary

The internet with its massive data–so called big data–allows researchers easily to extrapolate meaningful information out of an unstructured volume of spread news and comments. Text analysis algorithms search for specific mentions on Social Media and give investors an understanding of the interest of market participants. If company news or announcements were highly discussed on Social Media, we would assume that the stock price changes.

This research treats the question whether the information about the European economy posted on Twitter is related to the European stock market. The assumption is that the market does not distinguish between positive or negative information. Thus, the investigation focuses on a correlation analysis of the number of Tweets posted within a given time interval and the EURO STOXX 50, representing the European market. Further analysis focuses on the exploration of a forecasting model. With a linear regression, the author predicted the movement of intraday prices of the EURO STOXX 50, based on historical EURO STOXX 50 prices and the number of Tweets.

The data collection has been conducted for two types of variables: On the one hand, the author collected a financial data set of the EURO STOXX 50 index prices on a one-minute intraday interval and on the other hand a data set of the number of Tweets, mentioning the listed companies of the EURO STOXX 50 index. The statistical computing environment “R” offers an application programming interface to Twitter, which allowed to download Twitter data on a real time basis.

Between the observational periods from 13-19 September 2016 more than 370,000 Tweets have been collected. The number of Tweets in the morning was for all days lower than the number of Tweets in the afternoon. Hence, Twitter users are more active within the second half of the day.

It has been found that the number of Tweets posted within a one-minute time interval and the respective intraday price of the EURO STOXX 50 has a significant weak negative correlation. This result shows that when the number of Tweets is high, the variable “EURO STOXX 50” is rather low.

For the regression model the author analyzed the impact of the number of Tweets per minute to the movement of the EURO STOXX 50 price per minute. The linear regression analysis predicted a forecast model, which gives a forecast accuracy of the movement of the EURO STOXX 50 lying higher than 50%.

For further investigations and similar forecasting models the author recommends to analyze the parameter “good” or “bad” news within a text sentiment analyzing method. This helps to give an understanding of the quality of the information. In addition, it is recommended to observe data for a longer period of time to gain more accuracy.

VI. Appendix

EuroStoxx 50 companies

Name	Registered office	Industry
Air Liquide	Paris	Chemistry
Airbus Group	Amsterdam	Aerospace
Allianz	Munich	Insurance
Anheuser-Busch InBev	Belgium	Food and beverage
ASML Holding	Netherlands	Technology
Assicurazioni Generali	Trieste	Insurance
AXA	Paris	Insurance
Banco Bilbao Vizcaya Argentaria	Spain	Banking
Banco Santander Central Hispano	Spain	Banking
BASF	Germany	Chemistry
Bayer	Germany	Chemistry
BMW	Germany	Automotive industry
BNP Paribas	France	Banking
Carrefour	France	Retail
Compagnie de Saint-Gobain	France	Construction and materials
Daimler AG	Germany	Automotive industry
Deutsche Bank	Germany	Banking
Deutsche Post	Germany	Logistics
Deutsche Telekom	Germany	Telecommunication
E.ON	Germany	Electric utility
Enel	Italy	Electric utility
Engie	France	Electric utility
Eni	Italy	Petroleum
Essilor International	France	Pharmaceutical industry
Fresenius SE	Germany	Health care equipment
Groupe Danone	France	Food and beverage
Iberdrola	Spain	Electric utility
Inditex	Spain	Retail
ING Group NV	Netherlands	Insurance
Intesa Sanpaolo	Italy	Banking
L'Oréal	France	Personal and household goods

Name	Registered office	Industry
LVMH Moët Hennessy Louis Vuitton	France	Personal and household goods
Munich Re	Germany	Insurance
Nokia	Finland	Technology
Orange S.A.	France	Telecommunication
Philips	Netherlands	Personal and household goods
Repsol S.A. (excluded from index at 2015/09/01)	*Spain*	*Oil and gas*
Safran	France	Aerospace
Sanofi	France	Pharmaceutical industry
SAP SE	Germany	Technology
Schneider Electric	France	Goods and Services
Siemens	Germany	Goods and Services
Société Générale SA	France	Banking
Telefonica	Madrid	Telecommunication
TOTAL S.A.	France	Petroleum
Unibail-Rodamco	France	Real estate
UniCredit	Italy	Banking
Unilever	Netherlands/United Kingdom	Food and beverage
Vinci SA	France	Construction and materials
Vivendi	Paris	Media
Volkswagen Group	Germany	Automotive industry

Table 15 "Composition of EURO STOXX 50"; (STOXX Limited, 2016)

Statistics

Frequency of Tweets per day

Date	Total Number of Tweets [N_T]	Number of Tweets between 09:00-12:00 [N_M]	Number of Tweets between 12:01-17:31 [N_A]
13.09.2016	9230	5688	3542
14.09.2016	**76536****	8065*	**68471****
15.09.2016	36356*	**10580****	25776
20.09.2016	21106	6525*	14581
21.09.2016	22024	7733*	14291
26.09.2016	17986	3150	14836
27.09.2016	21192	6333*	14859
28.09.2016	28559	**9598****	18961
29.09.2016	16950	5020	11930
30.09.2016	8345	3123	5222
04.10.2016	26089	3097	22992
05.10.2016	2845	0	2845
06.10.2016	4932	0	4932
11.10.2016	15755	5426	10329
12.10.2016	16104	3918	12186
13.10.2016	15910	5745*	10165
14.10.2016	12100	4140	7960
19.10.2016	21231	6790*	14441
Grand Total	373250	94931	278319

* Value is higher than 2*sd (=$Outlier_1$)
** Value is higher than 3*sd (=$Outlier_2$)
Table 16 Number of Tweets per day

	Number of Tweets per day	*Morning*	*Afternoon*
Mean	20736	5274	15462
Standard Error	3823	673	3449
Median	17468	5557	13239
Standard Deviation	16218	2857	14633
Sample Variance	263031718	8163819	214125212
Kurtosis	8	0	11
Skewness	3	0	3
Range	73691	10580	65626
Minimum	2845	0	2845
Maximum	76536	10580	68471
Sum	373250	94931	278319
Count	18	18	18

Table 17 Descriptive statistics number of Tweets per day

Frequency of Tweets per minute ranged into bins

Lower range (number of Tweets per minute)	Upper range (number of Tweets per minute)	Frequency
0	10	2
11	20	132
21	30	1046
31	40	1892
41	50	1468
51	60	872
61	70	492
71	80	261
81	90	184
91	100	115
101	110	63
111	120	49
121	130	32
131	140	21
141	150	8
151	160	11
161	170	8
171	180	3
181	190	11
191	200	15
201	210	24
211	220	19
221	230	28
231	240	23
241	250	28
251	260	23
261	270	12
271	280	18
281	290	15
291	300	13
301	310	4
311	320	7
321	330	8
331	340	6
341	350	6
351	360	1
361	370	4
371	380	1
381	390	1
391	∞	0

Table 18 Frequency table for variable "Number of Tweets per minute"

Descriptive statistics variables per minute

Statistics

		EURO STOXX 50 (price per minute)	Number of Tweets per minute	Δ (abs change) EURO STOXX 50	Δ% (rel change) EURO STOXX 50
N	Valid	6926	6926	6925	6923
	Missing	0	0	1	3
Mean		2999,0000	53,891	,663096	,000220
95% Confidence Interval for mean					
	Lower bound		52,829		
	Upper bound		54,954		
Std. Error of Mean		,33562	,5420	,0101057	,0000032
Median		2999,2700	42,000	,449950	,000151
Mode		3019,54[a]	35,0	,0100	,0000
Std. Deviation		27,93117	45,1070	,8409587	,0002683
Variance		780,150	2034,646	,707	,000
Skewness		,135	3,711	9,943	9,252
Std. Error of Skewness		,029	,029	,029	,029
Kurtosis		-1,123	15,607	228,883	225,649
Std. Error of Kurtosis		,059	,059	,059	,059
Range		130,27	374,0	27,0400	,0091
Minimum		2932,55	6,0	,0000	,0000
Maximum		3062,82	380,0	27,0400	,0091
Sum		20771074,19	373250,0	4591,9369	1,5235
Percentiles	25	2974,8076	33,000	,199950	,000066
	50	2999,2700	42,000	,449950	,000151
	75	3021,1849	56,000	,880130	,000295

a. Multiple modes exist. The smallest value is shown

Table 19: Descriptive statistics

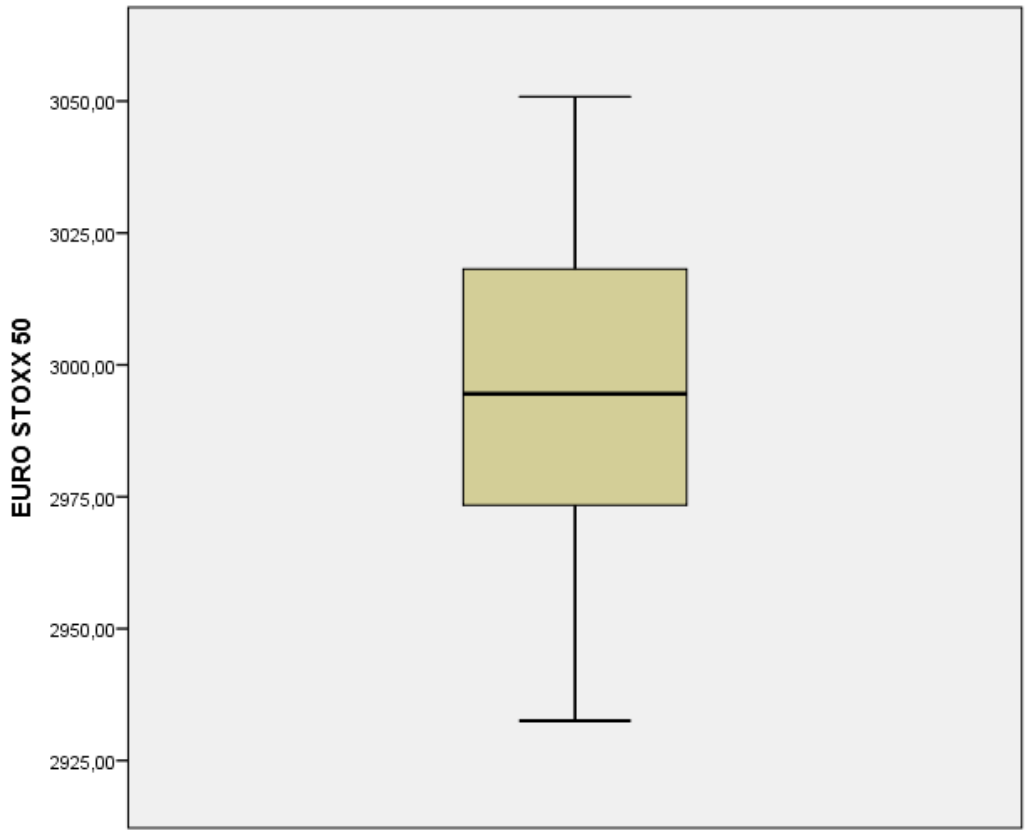


Figure 20 Boxplot of EURO STOXX 50

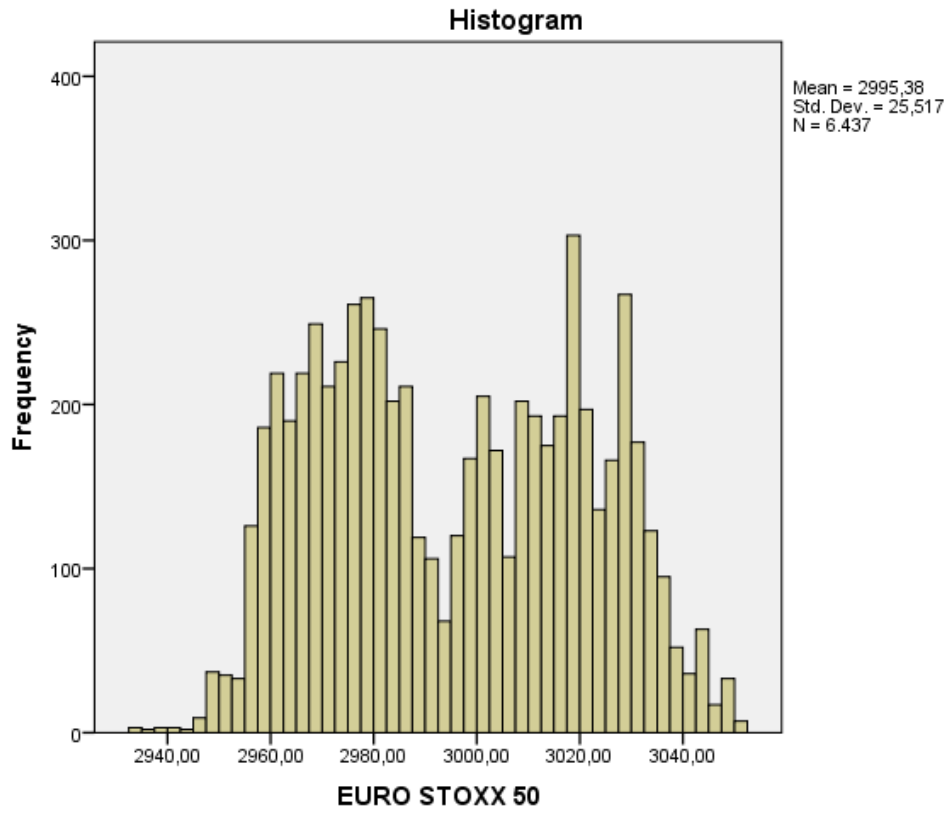


Figure 21 Histogram of the variable EURO STOXX 50

EURO STOXX 50 Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     3,00      293 .  &
     5,00      293 .  &
     5,00      294 .  &
    46,00      294 .  57888999&
    68,00      295 .  00012233444
   312,00      295 .  555556666666666777777777778888888888889999999999999
   409,00      296 .  000000000000000111111111111111222222222223333333333334444444444444
   468,00      296 .  555555555555566666666666666777777777777777888888888888888889999999999999999
   437,00      297 .  000000000000000011111111111112222222222222333333333333333344444444444444
   526,00      297 .  5555555555555556666666666666666666777777777777777888888888888888889999999999999999999
   448,00      298 .  000000000000000011111111111111122222222222222233333333334444444444444444
   330,00      298 .  55555555555555566666666666677777777777788888889999999
   174,00      299 .  000000111111112222223333444 4
   287,00      299 .  555555666666667777777777778888888889999999999999
   377,00      300 .  000000000000001111111111112222222222223333333333344444444444
   309,00      300 .  55555556666667777777777888888888888999999999999999
   368,00      301 .  000000000001111111111111122222222222233333333333444444444444
   496,00      301 .  555555555556666666666666677777777777778888888888888888999999999999999999999999
   333,00      302 .  000000000000000000011111111111222222233333333334444444
   433,00      302 .  555555556666666666667777777777777777888888888888888888899999999999999999
   300,00      303 .  0000000000000000011111122222222223333333344444444
   147,00      303 .  555555666666677777888999
    99,00      304 .  0112222233334444
    50,00      304 .  56788999
     7,00      305 .  0

 Stem width:     10,00
 Each leaf:       6 case(s)
```

Figure 22 Stem-and-Leaf Plot of the variable EURO STOXX 50

```
 Frequency    Stem &  Leaf

     2,00        0 .  &
    22,00        1 .  4&
   110,00        1 .  56778889999
   348,00        2 .  000001111111222222333333334444444 4
   698,00        2 .  555555555555666666666667777777777777788888888888888889999999999999999
   916,00        3 .  000000000000000011111111111111111122222222222222223333333333333333334444444444444444444444
   976,00        3 .  555555555555555555556666666666666666667777777777777777778888888888888888889999999999999999
   825,00        4 .  000000000000000000111111111111111111222222222222222233333333333333334444444444444444
   643,00        4 .  55555555555555666666666666666677777777777788888888888889999999999
   491,00        5 .  000000000000111111111222222222223333333334444444
   381,00        5 .  555555555666666667777777788888888999999
   262,00        6 .  000000011111112223333344 44
   230,00        6 .  5555566666777788889999
   155,00        7 .  000111222233344
   106,00        7 .  55566778899
   104,00        8 .  00011233344
    80,00        8 .  5567889
    17,00        9 .  00
   560,00 Extremes    (>=91)

 Stem width:      10,0
 Each leaf:      10 case(s)
```

Figure 23 Stem&Leaf Plot of the variable number of Tweets per minute

Normality discussion of transformed variable “Number of Tweets per minute”

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Number of Tweets	6926	6,0	380,0	53,891	45,1070	3,711	,029	15,607	,059
Log	6926	,78	2,58	1,6551	,22788	1,248	,029	2,631	,059
Squareroot	6926	2,45	19,49	6,9880	2,24943	2,501	,029	7,799	,059
Divided	6926	,00	,17	,0248	,01110	1,339	,029	9,365	,059
Valid N	6926								

Table 20 Descriptive statistics of transformed variables

Correlations

		EURO STOXX 50	Δ (abs change)	Δ% (rel change)	Number of Tweets
EURO STOXX 50	Pearson Correlation	1	-,105**	-,117**	-,241**
	Sig. (2-tailed)		,000	,000	,000
	Sum of Squares and Cross-products	5402541,017	-17077,729	-6,082	-2103427,995
	Covariance	780,150	-2,466	-,001	-303,744
	N	6926	6925	6923	6926
Δ (abs change)	Pearson Correlation	-,105**	1	1,000**	,043**
	Sig. (2-tailed)	,000		,000	,000
	Sum of Squares and Cross-products	-17077,729	4896,733	1,486	11422,306
	Covariance	-2,466	,707	,000	1,650
	N	6925	6925	6923	6925
Δ% (rel change)	Pearson Correlation	-,117**	1,000**	1	,051**
	Sig. (2-tailed)	,000	,000		,000
	Sum of Squares and Cross-products	-6,082	1,486	,000	4,261
	Covariance	-,001	,000	,000	,001
	N	6923	6923	6923	6923
Number of Tweets	Pearson Correlation	-,241**	,043**	,051**	1
	Sig. (2-tailed)	,000	,000	,000	
	Sum of Squares and Cross-products	-2103427,995	11422,306	4,261	14089921,916

	Covariance	-303,744	1,650	,001	2034,646
	N	6926	6925	6923	6926

**. Correlation is significant at the 0.01 level (2-tailed).

Table 21 Correlation between variables

Regression analysis:

Model 1:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	EURO STOXX t-1, Number of Tweets t-1[b]	.	Enter

a. Dependent Variable: EURO STOXX 50

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,998[a]	,996	,996	1,69068

a. Predictors: (Constant), EURO STOXX t-1, Number of Tweets t-1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4172081,608	2	2086040,804	729794,381	,000[b]
	Residual	18390,915	6434	2,858		
	Total	4190472,523	6436			

a. Dependent Variable: EURO STOXX 50

b. Predictors: (Constant), EURO STOXX t-1, Number of Tweets t-1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4,611	2,556		1,804	,071
	Number of Tweets t-1	,001	,000	,003	2,999	,003
	EURO STOXX t-1	,998	,001	,998	1172,876	,000

a. Dependent Variable: EURO STOXX 50

Model 2:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
2	EURO STOXX t-1[b]	.	Enter

a. Dependent Variable: EURO STOXX 50

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2	,998[a]	,996	,996	1,69173

a. Predictors: (Constant), EURO STOXX t-1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	4172055,898	1	4172055,898	1457768,668	,000[b]
	Residual	18416,626	6435	2,862		
	Total	4190472,523	6436			

a. Dependent Variable: EURO STOXX 50

b. Predictors: (Constant), EURO STOXX t-1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
2	(Constant)	6,540	2,476		2,642	,008
	EURO STOXX t-1	,998	,001	,998	1207,381	,000

a. Dependent Variable: EURO STOXX 50

Model 3:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
3	Number of Tweets t-1[b]	.	Enter

a. Dependent Variable: EURO STOXX 50

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3	,239[a]	,057	,057	24,77720

a. Predictors: (Constant), Number of Tweets t-1

b. Dependent Variable: EURO STOXX 50

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
3	Regression	239964,242	1	239964,242	390,879	,000[b]
	Residual	3950508,282	6435	613,910		
	Total	4190472,523	6436			

a. Dependent Variable: EURO STOXX 50

b. Predictors: (Constant), Number of Tweets t-1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
3	(Constant)	3002,548	,476		6303,299	,000
	Number of Tweets t-1	-,131	,007	-,239	-19,771	,000

a. Dependent Variable: EURO STOXX 50

Program codes

Installation of Twitter environment for "R" and connection to Twitter API:

```
install.packages('streamR')
install.packages('ROAuth')
library(ROAuth)
library(streamR)

credential <- OAuthFactory$new(consumerKey='xxxxxxxxxxxxxxxxxxxxxx',
consumerSecret='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx',
        requestURL='https://api.twitter.com/oauth/request_token',
        accessURL='https://api.twitter.com/oauth/access_token',
        authURL='https://api.twitter.com/oauth/authorize')

options(RCurlOptions = list(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl")))
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

credential$handshake(cainfo="cacert.pem")
```

Filter request to API:

```
filterStream( file.name="tweets_eurostoxx_0210.json",
        track= "$ABI.BR, $AI.PA, $AIR.PA, $ALV.DE, $ASML.AS, $BAS.DE, $BAYN.DE, $BBVA.MC, $BMW.DE, $BN.PA, $BNP.PA, $CA.PA, $CS.PA, $DAI.DE, $DBK.DE, $DG.PA, $DPW.DE, $DTE.DE, $EI.PA, $ENEL.MI, $ENI.MI, $EOAN.DE, $FP.PA, $G.MI, $GLE.PA, $GSZ.PA, $IBE.MC, $INGA.AS, $ISP.MI, $ITX.MC, $MC.PA, $MUV2.DE, $NOK1V.HE, $OR.PA, $ORA.PA, $PHIA.AS, $REP.MC, $RWE.DE, $SAN.MC, $SAN.PA, $SAP.DE, $SGO.PA, $SIE.DE, $SU.PA, $TEF.MC, $UCG.MI, $UL.PA, $UNA.AS, $VIV.PA, $VOW3.DE, ABI.BR,Anheuser-Busch InBev SA/NV, AI.PA, L'Air Liquide SA, AIR.PA, AIRBUS GROUP, ALV.DE, Allianz SE, ASML.AS, ASML HLDG, BAS.DE, BASF SE, BAYN.DE, Bayer AG, BBVA.MC, BBVA, BMW.DE, Bayerische Motoren Werke Aktiengesellschaft, BN.PA, Danone, BNP.PA, BNP Paribas SA, CA.PA, Carrefour SA, CS.PA, AXA Group, DAI.DE, Daimler AG, DBK.DE, DeutscheBank AG, DG.PA, VINCI S.A., DPW.DE, DeutschePost AG, DTE.DE, Deutsche Telekom AG, EI.PA, Essilor International SA, ENEL.MI, Enel SpA, ENI.MI, Eni SpA, EOAN.DE, E.ON SE, FP.PA, TOTAL S.A., G.MI, Assicurazioni Generali S.p.A., GLE.PA, Societe Generale Group, GSZ.PA, ENGIE SA, IBE.MC, IBERDROLA, INGA.AS, ING GROUP, ISP.MI, Intesa Sanpaolo S.p.A., ITX.MC, INDITEX, MC.PA, LVMH Moët Hennessy Louis Vuitton SA, MUV2.DE, Münchener Rückversicherungs-Gesellschaft Aktiengesellschaft, NOK1V.HE, Nokia Corporation, OR.PA, L'Oreal SA, ORA.PA, PHIA.AS, ROY.PHILIPS, REP.MC, REPSOL, RWE.DE, RWE AG, SAN.MC, BANCO SANTANDER, SAN.PA, Sanofi, SAP.DE, SAP SE, SGO.PA, Compagnie de Saint-Gobain S.A., SIE.DE, Siemens Aktiengesellschaft, SU.PA, Schneider Electric SE, TEF.MC, TELEFONICA, UCG.MI, UniCredit
```

S.p.A., UL.PA, UNIBAIL-RODAMCO, UNA.AS, UNILEVER CERT, VIV.PA, Vivendi S.A., VOW3.DE, Volkswagen AG", oauth=credential, timeout=43200, verbose=TRUE)

Text counting analysis (Anderson, 2017)**:**

```
temple.text<-scan(choose.files(),what="char",sep="\n")
temple.text<-tolower(temple.text)
temple.words.list<-strsplit(temple.text,"\\W+",perl=TRUE)
temple.words.vector<-unlist(temple.words.list)
temple.freq.list<-table(temple.words.vector)
temple.sorted.freq.list<-sort(temple.freq.list,decreasing=TRUE)
temple.sorted.table<-paste(names(temple.sorted.freq.list),temple.sorted.freq.list,sep="\t")
cat("Word\tFREQ",temple.sorted.table,file=choose.files(),sep="\n")
```

References

Anderson, J. V. (2017, 03 08). *John Victor Anderson dot org*. Retrieved from http://johnvictoranderson.org/?p=115

Antweiler, W. & Murray, F. Z. (2005). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance, 59*, pp. 1259–1294.

Barbera, P. (2014, January 7). *Access to Twitter Streaming API via R.* Retrieved May 19, 2016, from https://www.r-project.org/

Bayer AG. (2017, June 06). *Investor News 2016*. Retrieved from http://www.investor.bayer.de/en/nc/news/archive/investor-news-2016/investor-news-2016/?tx_news_pi1%5Bnews%5D=1925&cHash=06f641f0290d3db2b3bc73ab9f56ace0

Boyd, D. & Ellison, N. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication, 13 (1)*, pp. 210-230.

Brannath, W., Futschik, A. & Krall, C. (2010). *Statistik im Studium der Wirtschaftswissenschaften* (3. Auflage). Wien: Facultas Verlags- und Buchhandels AG.

Byrne, A., & Brooks, M. (2008). Behavioral Finance: Theories and Evidence. *The Research Foundation of CFA Institute*, pp. 1-26.

Crocoll, S. (2012, June 06). "*Twitter weiß es besser*". Retrieved from www.zeit.de: http://www.zeit.de/2012/24/F-Soziale-Netzwerke/komplettansicht

Danneman, N., & Heimann, R. (2014). *Social Media Mining with R* . Birmingham: Packt Publishing Ltd.

Die Zeit/Wirtschaft. (2017, June 10). *www.zeit.de*. Retrieved from http://www.zeit.de/wirtschaft/2016-09/deutsche-bank-rettungsplan-finanzaufsichtsbehoerde

Fama, E. (1965). The Behavior of Stock Market Prices. *Journal of Business, 38*, pp. 34-105.

Fama, E. (1970b). Efficient capital markets: A review of theroy and empirical work. *Journal of Finance, 25*, pp. 383-417.

Fama, E. (1991). Efficient Capital Markets: II. *The Journal of Finance, 46*, pp 1575-1617.

Field, A. (2005). *Discovering Statistics Using SPSS* (5th ed.). London: SAGE Publications Ltd.

Goldberg, J. (2014, April 26). *Wallstreet Online*. Retrieved from https://www.wallstreet-online.de: https://www.wallstreet-online.de/nachricht/5086288-behavioral-finance-economics-geruechtekueche

Grossman, S. J. & Stiglitz, J. E. (1981). The determination of the variability of stock market prices. *American Economic Review, 71*, pp. 222-227.

Hoyer, N. (2011, January 13). *Handelsblatt*. Retrieved from www.handelsblatt.com: http://www.handelsblatt.com/panorama/aus-aller-welt/verdacht-auf-kursmanipulation-rapper-50-cent-treibt-kurs-der-eigenen-aktie-ueber-twitter/3762458.html

Investopedia. (2017, May 28). *investopedia.com*. Retrieved from http://www.investopedia.com/terms/v/volatility.asp

Jensen, M. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economic Review, 6*, pp. 95-101.

Makhabel, B. (2015). *Learning Data Mining with R*. Birmingham: Packt Publishing Ltd.

Malkiel, B. (1989). *Is the Stock Market Efficient*? Science, Vol. 243, No. 4896, pp. 1313-1318.

Malkiel, B. G. (1973). *A Random Walk Down Wall Street.* New York: W. W. Norton & Company.

Mao, Y. (2012). Correlating S&P 500 Stocks with Twitter Data. *Hot Social*, 4. edition

Prajapati, V. (2013). *Big Data Analytics with R and Hadoop*. Birmingham: Packt Publishing Ltd.

Sakaki, T., Okazaki, M. & Matsuo, Y. (2010). Earthquake Shake Twitter users: Real-time Event Detection by Social Sensors. *WWW*.

Samuelson, P. (1965). Proof that Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review, 6*, pp. 41-49.

Schumaker, R. P. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems, 27 (2)*, pp. 1-19.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66 (3)*, pp. 605-610.

Shumway, R. & Stoffer, D. (2011). *Time Series Analysis and Its Applications* (3rd ed.). New York: Springer Science+Business Media, LLC.

STOXX Limited. (2016). *EURO STOXX 50*. Retrieved May 19, 2016, from https://www.stoxx.com/index-details?symbol=SX5E

Tayal, D. & Komaragiri, S. (2009). Comparative Analysis of the Impfact of Blogging and Micro-blogging on Market Performance. *International Journal, 1 (3)*, pp. 176-182.

The R foundation. (1993). *The R Project for Statistical Computing*. Retrieved May 18, 2016, from https://www.r-project.org

Tversky, A. & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and Biases. *Science*, pp. 1124-1131.

Wikipedia (2017). *Wikipedia, Brexit*. Retrieved 01 12, 2017, from https://en.wikipedia.org/wiki/Brexit

Wikipedia (2017). *Wikipedia, Volkswagen emissions scandal*. Retrieved 01 12, 2017, from https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal

Wolfram, M. S. (2010). Modelling the Stock Market using Twitter. *University of Edingburgh*, p. 65.

Wysocki, P. D. (1991). Cheap Talk on the Web: The Determination of Postings on Stock Message Boards. *University of Michigan*(Working Paper).