CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Vladimír Zobal

Sentiment analysis of social media and its relation to stock market

Bachelor thesis

Prague 2017

Author: Vladimír Zobal

Supervisor: doc. PhDr. Ladislav Krištoufek, Ph.D.

Academic Year: 2016/2017

Number of characters (with spaces): 58 777

Bibliographic Evidence Card

ZOBAL, Vladimír. *Sentiment analysis of social media and its relation to stock market.* Prague 2017. 55 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor: doc. PhDr. Ladislav Krištoufek Ph.D.

Abstract

The aim of this thesis is to verify and quantify relationship between the sentiment of Twitter posts, so called Tweets, and shot-run movements of the stock market indices. Firstly, the sentiment analysis, using lexicon-based approach, is conducted and subsequently predictions of stock market indices for the next day are made. Our analysis focuses mainly on noise reduction by implementation of threshold, and also on differences between negative and positive sentiment in contrast with reaction of individual market actors. All predictions are conducted on daily basis and the main purpose of this model is to support trading decision of individual by inclusion of sentiment score obtained from Twitter.

Keywords	Sentiment analysis, Social media, Stock market, Twitter, Tweets, Trading, Financial forecasting, Decision-supporting model, Behavioral economics

Author's e-mail	**vladimir.zobal@gmail.com**
Supervisor's e-mail	**ladislav.kristoufek@fsv.cuni.cz**

Bibliografický záznam

ZOBAL, Vladimír. *Sentiment analysis of social media and its relation to stock market.* Praha 2017. 55 stran. Bakalářská práce (Bc.) Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií. Vedoucí práce: doc. PhDr. Ladislav Krištoufek Ph.D.

Abstrakt

Tato bakalářská práce si stanovila za cíl ověřit a kvantifikovat vztah mezi sentimentem příspěvků na sociálni sítí Twitter a s jeho pomocí objastnit krátkodobý vývoj ukazatelů akciového trhu. Nejprve je provedena sentimentová analýza založená na přístupu, využívajícím slovník obsahujíci informaci o sentimentu jednotlivých výrazů a následně jsou předpovídány klíčové akciové ukazatele pro následující den. Hlavním cílem naší analýzy je odstranění šumu zavedením hranice, od které je sentiment vzatý v potaz. Dílčím cílem je analýza rozdílu mezi reakcí jednotlivých ekonomických aktérů pod vlivem odděleného negativního a pozitivního sentimentu.

Všechny předpovědi jsou prováděny na denní bázi a hlavním účelem tohoto modelu je pomoc při rozhodování ohledně investice do akcií s využitím sentimentového hodnocení sociální sítě Twitter.

Klíčová slova	Sentimentová analýza, Sociání média, Akciový trh, Twitter, Trading, Finanční předpovídání, Behaviorální ekonomie

E-mail autora	**vladimir.zobal@gmail.com**
E-mail vedoucího	**ladislav.kristoufek@fsv.cuni.cz**

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 31 July 2017

Signature

Acknowledgment

I would like to express my sincere gratitute to my supervisor, doc. PhDr. Ladislav Krištoufek Ph.D., for his remarks and patient guidance, without which it would not be possible to complete this thesis.

Bachelor Thesis Proposal

Author	Vladimír Zobal
Supervisor	doc. PhDr. Ladislav Krištoufek Ph.D.
Proposed topic	Sentiment analysis of social media and its relation to stock market

---

Topic characteristics and hypothesis

The aim of this thesis is to enrich selected financial forecasting models with an additional factor - actual sentiment on social networks. For the purpose of this paper we decided to focus on Twitter only. Hypothesis for verification: If the sentiment of Twitter posts, known also as tweets, is mostly positive, the stock price of a mentioned company should be growing and vice versa. Technical part of the thesis will be done using statistical language R, the most important part is pre-preparation of data for analysis, subsequent evaluation of sentiment of each tweet and interconnection of this variable with selected already existing models.

Overall outcome of the thesis should be applicable model able to predict future behavior of target stock more precisely, than model without this variable. Necessary prerequisites for an efficient application of this model are relatively high interest in selected stock during last few days causing noise reduction and unambiguously evaluated sentiment.

Leading challenge is analysis of tweets, which have their length limited to 140 characters and also there are quite often used special words, which is significant complication for direct text mining.

Outline

1. Introduction
2. Literature review
3. Methodology

4. Results and interpretation
5. Conclusion

Core bibliography

1. SALAS-ZÁRATE, María del Pilar, José MEDINA-MOREIRA, Paul Javier ÁLVAREZ-SAGUBAY, Katty LAGOS-ORTIZ, Mario Andrés PAREDES-VALVERDE a Rafael VALENCIA-GARCÍA. Sentiment Analysis and Trend Detection in Twitter [online]. s. 63 [cit. 2017-01-30]. DOI: 10.1007/978-3-319-48024-4_6. Available from: http://link.springer.com/10.1007/978-3-319-48024-4_6

2. WOLFRAM, M. S. A. (2010). Modelling the Stock Market using Twitter. Master thesis, The University of Edinburgh.

3. SHLEIFER, A. (2000). Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press.

4. O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. R., and SMITH, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM, 11(122-129):1-2.

5. GROLEMUNG, G. (2014): Hands-On Programming with R, O'Reilly Media Inc., Sebastopol, CA, USA

6. ZUMEL, N. & MOUNT, J. (2014): Practical Data Science with R, Manning Publications Co., Shelter Island, NY, USA.

Contents

Acronyms

API Application Programming Interface
BOW Bag-Of-Words
CL Computational Linguistic
DJIA Dow Jones Industrial Average
DL Deep Learning
DM Diebold-Mariano test
DTM Document Term Matrix
EMH Efficient Market Hypothesis
ET Eastern Time zone
GDP Gross Domestic Product
GPOMS Google-Profile Of Mood States
IDF Inverse Document Frequency
KPI Key Performance Indicator
LMI Lexicographer's Mutual Information
MPQA Multi Perspective Question Answering
NLP Natural Language Processing
NYSE New York Stock Exchange
OLS Ordinary Least Squares
RWH Random Walk Hypothesis
SA Sentiment Analysis
S&P 500 Standard & Poor's 500 index
SVM Support Vector Machines
TA Text Analytics
TSA Twitter Sentiment Analysis
URL Uniform Resource Locator

1 Introduction

The stock market is a place, where investors could either make a fortune or lose absolutely everything they possess and sometimes even more. Various strategies were developed with the intention to beat the market. Their results differ widely, yet overall profits should correspond to the outcome of zero-sum game, i. e. that the total sum of winnings and losses of various market players is always equal to zero. There is no way to create a perfect model, because there are too many variables, we are just not able to account for. But with a little exaggeration we can say that every model which in predicting outperforms a coin flip is a success. With the advent of the data science, new and complex trading strategies based on Natural Language Processing (NLP) were created. Large body of literature has also been devoted to this topic and during recent years quite a few of them combine traditional approaches with data mining to dig the patterns from ex-post data and use these patterns to predict future behaviour based on historical values.

The aim of this thesis is to prove correlation between Social media sentiment and Stock market returns. This idea is based on approach of behavioural economics, which profoundly links emotions with individual decision making and behaviour. It explicitly means, besides other things, that we would like to reject Efficient Market Hypothesis (EMH). According to the EMH, stocks are always traded at their fair value on stock exchanges, making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices, because existing share prices always incorporate and reflect all relevant information immediately at the time they became publicly known (Investopedia, 2017). We suppose that changes are fully reflected in the stock prices with some kind of a lag. As a result of such lag, space for buying under-valuated stocks or selling currently over-valuated ones is created.

We used all mentioned approaches to evaluate Twitter sentiment and tried to determine a speculative strategy based on the result of Twitter

Sentiment Analysis (TSA) on real data. The key aspect of that strategy is to obtain sufficient information early enough to make a decision ahead of other investors and gain an advantage in investment opportunities. This might be quite challenging task considering the speed of spreading information nowadays. We do not intend to clarify all sudden price changes - these are extremely hard to predict due to an impact of stock market actors like investment funds, "big players" or inside traders, yet we are able to cover movements caused by other kinds of investment.

The uniqueness of this thesis lies in combination of various techniques into a single model and limitation of estimation only to periods with abnormal activity. For the purpose of reducing the noise, we added a activity-based threshold into our model. The aim of this model is not to give the most precise estimate, but rather to indicate noticeable changes in price prior to adaption of other investors and market itself. The strictest adaptation window for intra-day predictions was estimated firstly by Gidofalvi to 20 minutes after the financial news article is published (2001) and then even reduced by Fortuny et al. to 8 minutes, in which information should be fully contained in actual market price (2013). We propose following acceptance criterion to the model performance be counted as success. First of them is processing time below 15 minutes with accuracy at least 65 % together with the goal of achieving better returns than human traders by removing the elements of emotion and bias from trading (Jelveh, 2006).

We do not strictly want to create a decision making tool itself; we are rather trying to develop decision supporting tool, which would be able to identify profitable investment opportunities from pre-selected list of companies of interest. Target trading profile for such strategy is the conservative one. Essential step prior modelling itself is creation of subset of companies, where investor is willing to invest, however he is waiting for the most suitable time to do so. In the contrary to EMH, we assume that people do not always act rationally even if they have enough relevant information, they are trying to avoid losses and when they are starting to lose money, they often tend to overreact - sell, when other participants do so, with only reason to minimize short-run loss. On the other hand, we suppose that long lasting positive sentiment might lead to overvaluation and even to market bubbles, while global pessimism, makes stocks undervalued, which might open space for purchase at advantageous price (Ni, 2015).

To the best of our knowledge, there has been no research conducted yet using thresholds placed on Twitter sentiment score resulting in inclusion or exclusion of sentiment parameter in the model. Inclusion of this feature, as well as utilization of various sentiment analysis techniques, should be considered as major contribution of this thesis.

This bachelor thesis is organized in the following manner: The first chapter outlines the background, motivation and organization of the thesis. The second chapter summarizes previous researches regarding the topic and corresponding subtopics, starting with its history and in the end refers to the most current trends and researches conducted in the field. The third chapter is devoted to methodology and evaluation of available pre-processing techniques and algorithms for evaluation of sentiment of micro-blogging data, where pros and potential cons of each alternative are elevated. The fourth chapter describes the structure of model for prediction of future price movements, which incorporates factor obtained thanks to the third part. The last chapter of theoretical part is devoted to data used for analysis with detailed workflow and preliminary processes. Following empirical section, is the most important part of this thesis - it introduces results of predictive model and also evaluates research questions proposed within section covering methodology. The thesis ends with conclusion, where possible future tasks and improvements are suggested.

2 Literature Review

2.1 Stock Market Modeling

In the very beginning, everything started with two major theories that have had an impact on market predictions, Fama's Efficient Market Hypothesis (EMH) and Malkiel's Random Walk Hypothesis (RWH). Both of them reject possibility of predicting future price movements with accuracy higher than 50 %. EMH claims that, all available information is already completely reflected in current prices and agents have rational expectations. It needs to be mentioned, that there is no limitation on adequacy of reaction of an individual investor, who might overreact or underreact. Lately, this theory was extended to three common forms in which the hypothesis is now stated: weak, semi-strong and strong. Week form denies possibility to predict future price by analysis of the past performance patterns, which means that patterns in past will not occur anymore. Generally it only means that agents are not able to profit from inefficiencies in the long run, since this inefficiencies should be more or less random. Furthermore Semi-strong version also reflects all new publicly known information in an unbiased fashion very fast and eventually strong version treats the same way as semi-strong version also insider information.

The second hypothesis mentioned, RWH, suggests that the "future path of the price level of a security is no more predictable than the path of a series of accumulated random numbers" (Malkiel, 1999). This means that ex-ante estimates of future behaviour seem to be pointless because changes in stock prices should behave like i.i.d random variables not correlated with any information available. This also places limitations on possible accuracy, already mentioned, which cannot be more than 50 %.

Stock market can be perceived as a collective decision making system, influenced externally by public opinion as well as internally by performance of individual stocks (Zhang et al., 2016). It would be childish to suppose that all investors are perfectly aware of every piece of news at the same time, more precisely at the time information becomes publicly known or sometimes before relevant cause even happens. Even if someone would have resources to collect all available information at time they are published, it would be extremely hard and demanding to evaluate all inputs in time correctly. The value of lag, between publishing of news and it's incorporation in the stock price, proposed by Lebaron was lately estimated to 20 minutes, which is not a long time considering that investors do not know what piece of information are they looking for and even if something to be looked for exists at all.

Factors influencing a performance of stock market can be divided into following three groups: Investor sentiment indicators, Business-cycle variables and factors related to the financial market (Kadili, 2015). We have devoted separate section to investor sentiment indicator, so here we look cursorily on the remaining two groups. Without any doubt, phase of business cycle, that can be captured by factors as growth of real GDP, inflation, short-term interest rates and term spread between long-term and short-term government bond yields, influence return of stocks and bonds significantly. The second group contains factors related directly to the stock market, such as volatility, dividend yield, liquidity and market size. Each of variables mentioned in this paragraph has an impact on the stock market, nevertheless we will not be discussing their influence in detail here, in our model this part is treated as a black box substituted by growth of corresponding index, which should absorb all effects of these variables on its own.

2.2 Sentiment Analysis

2.2.1 Traditional Sentiment Analysis

Close connection between mood, emotions and decision-making is obvious - person with a good mood thinks about everything more positively and vice versa. Positive mood can also lead to underestimation of possible risks due to the main focus on positive consequences of individual decisions (Loewenstein, 2010) and also inclination of people to spend less time with making a decision, while ignoring information they might otherwise find relevant (Isen

& Means, 1983). In the presence of uncertainty, overall effect of sentiment on decision-making tends to be even stronger (Porshnev et al., 2016), which is without any doubt true in case of stock-market trading (Lepori, 2015). These are like a foundation stones of our work, because existence of such correlation is essential prerequisite for our model.

Sentiment analysis (SA), also known as opinion mining, is a type of data mining that measures the inclination of people's opinions through natural language processing (NLP), computational linguistics (CL) and text analysis (TA), which are used to extract and analyze subjective information from the Internet. History of sentiment analysis dates back to the beginning of 21st century, when Pang et al. wrote one of the first papers regarding the topic. During the research 69 % accuracy was obtained with the list of seven positive and seven negative words used for the training their model (Pang et al., 2012). Most of the former papers still follows their work, building a classifier based on annotated corpus with manually-designed sophisticated features (Ren et al., 2016). Untill the 2013, most of the studies were devoted to analysis of news articles, annual financial reports and also other financial-related information available on the Internet. Among others they analyzed a sentiment of movie reviews with the use of unsupervised machine learning techniques, because apart from written review, IMDb also collects corresponding star rating that was used for training of the model (Kearney & Liu, 2014).

Most of the primary models based on unsupervised techniques is using a lexicon of opinion words with sentiment score for sentiment evaluation. Original lexicons contained only list of single words and their sentiment, the most commonly used were: *AFFIN* (Nielsen, 2011), *BING*, *MPQA*, *SentiWordNet* or Mohammad et al.'s *NRC* (2013), which were mostly suitable for general-purpose sentiment analysis. The only exception from lexicons listed above is made by the latter most, which was designed specifically for a use on Twitter. Each of them uses different metric for sentiment evaluation, whereas *AFFIN* assigns scores that runs between -5 and 5 to indicate sentiment of the word, NRC assigns yes or no to each of researched emotion, which runs from casual positive and negative to anger, anticipation, disgust, fear etc. The last one to mention, BING, classifies words in binary fashion (positive/negative). What might also cause troubles is the fact, that these lexicons are not balanced - they contain significantly more negative words than positive. Most of these lexicons were validated for general-purpose and we found them inappropriate for analysis of micro-blog financial content. Beside other things, these methods do not take into account different sentiment for negations, sarcasm and also are not able to cope with acronyms.

"However, even customized platform-specific lexicons still suffer from ambiguities at a contextual level, e.g. cold beer (+) or cold food (-), dark chocolate (+) or dark soul (-)" (Flekova et al., n/a), which demonstrates problem of different sense of one word even on word-level. According to Cook & Stevenson (2010) another rather problematic aspect of this approach is instability of lexicons across not only domain, but also time, these changes were deeper analyzed by Mitra et al. (2014). Countless number of authors had been dealing with settle up with negations (Moilanen & Pulman, 2008); (Choi & Cardie, 2009) and polar words, which can carry in new domain even absolutely opposing sentiment (Schwartz et al., 2013). This especially holds in terms of general-purpose versus financial-purpose. Machine learning approach detecting shifted polarity was proposed by Ikeda et al. (2008). Despite extensive work on polarity modifiers, especially intensifiers, we did not find any evidence of algorithm that can cope with differences between *cannot be bad* and *cannot be worse* (Flekova et al., *n/a*).

In the contrary to lexicon based methods, another unsupervised method is bag-of-words (BOW), which is a corpus created from manually annotated Tweets based on inclusion of selected feature and frequency of selected number of top terms (Magdy et al., 2015). Subsequently testing set of Tweet is labeled with the use of minimization of fitting error by Support Vector Machine (SVM) algorithm.

Another fields that were found exploitable by sentiment analysis of micro-blogging and web content were retail sales and services, band popularity (Vries et al., 2012), politics (Partelová, 2017) and also disease breaks as a result of analysis of Google Queries (Tu et al., 2015). The last mentioned section is really interesting, Eiji et al. have studied data from Twitter and Google trends to make early predictions of disease outbreaks (2011).

The sentiment classification task can be handled by several approaches. The most commonly used ones are Lexicon-based approach, which is also the most prevalent thanks to its relative simplicity, Machine-learning approach and hybrid modelings. Probably the most accurate, but extremely expensive method is the use of manual annotation. This can be only provided with at least 3 independent annotators, who know the context and are able to quickly classify enough data to make prediction.

2.2.2 Twitter and StockTwits sentiment Analysis (TSA)

Sentiment analysis of micro-blogging statuses differs from traditional sentiment analysis in various aspects. The most important one is the length of messages and also frequency of non-standard words used, because traditional SA aimed especially on longer pieces of texts (reviews, articles etc.). On the other hand, these messages give preference to content and mass scale towards to form. One of the modern and extremely fast source of news and opinions is social network called Twitter, launched in 2006, with more than 313 million active users worldwide on monthly basis. Users interact with each other via public messages, "Tweets," restricted to only 140 characters. Tweets are mostly written in spoken linguistic style of more than 40 national languages and also include special expressions ranging from emoticons, hashtags (#) , cash-tags ($) and hyperlinks to internet slang. Tweets can also contain another user generated content. They can be liked, shared by the community. Information about number of those interaction is also recorded along with a wording of message, which can be used for elimination of noise (For details see appendix A). Due to the large user base, accessibility and relative informality valuable information can be obtained extremely fast.

One of the first models, conducted by Go et al., was only able to distinguish between ”positive” and ”negative” sentiment (2009), lately Pak et al. coded classifier to detect also another group named ”neutral” (2010), which is quite useful for handling Tweets containing no sentiment information. It is rational to suppose that majority of Tweets do not contain unambiguous sentiment or they are only factual and contain no sentiment at all. These models required quite large set of manually annotated Tweets for training of the model, which made them quite expensive and time-demanding. One of the first extensive lexicon with annotated Tweets corpus was created in contest SemEval-2013 by Nakov et al. (2013) and contained contextual phrase-level as well as polarity score.

Plenty of data-scientist based their research on special features of Twitter messages. Liebrecht et al. (2013) focused on hashtag *#sarcasm*, which should result in flipped polarity of a Tweet. Issues connected with sarcasm identification were also in detail explored by Gonzalez-Ibanez et al. (2011).

Another approach is to manually evaluate large data set of Tweets and than use supervised machine learning techniques, but this is rather expensive and time-consuming procedure.

The most sophisticated contemporary models are based on deep learning algorithms (DL), which are able to *discover multiple levels of representation, with higher-level features representing more abstract aspects of the data.* Deep learning belongs to the machine learning section of computer science, while the main difference between Deep learning and other methods lies in specification of what patterns the model should be looking for. DL has an ability to extract and organize the discriminating information from the data and is less dependent on feature engineering (Bengio, 2013). Alternatively to this approach, Magdy et al. (2015) classified Tweets using distant supervision. This approach does not only take into account text of Tweet alone, but it also attaches category and rating connected to linked video if there is YouTube link included. Topic information obtained from server *YouTube.com* can serve well for creation of topic specific classifier, which outperforms general classifiers in accuracy as well as robustness, because there is no need for manually annotated data and thanks to that updates can be than provided more often without high additional costs (Magdy & Elsayed, 2014).

2.3 Sentiment Analysis used for Predicting Stock Market

One might think that stock trading is just about financial and technical data and their analysis, but we need to take into consideration other quite important factors: expectations of investors and their faith in the future of the company they are interested to invest in, which can be exaggeratedly influenced by the news and might intensify overall effect of any news. That is when Sentiment analysis comes to play its role. Stock price can be appreciated as sum of fundamental value influenced by rational investors and risk premium inflected by noise traders (Ni, 2015). Typical feature for the latest is overreaction, which is a result of cognitive bias of overconfidence - constant overestimation of accuracy of own judgments (Pallier et al., 2012). Investor sentiment influences stock valuation and can cause biased expectations, such as the propensity to speculate and investor's optimism or pessimism on stock real valuation (Isen & Means, 1983).

According to Google-Profile of mood states (GPOMS), mood can be classified into one or more of 6 dimensions, Calm, Alert, Sure, Vital, Kind, and Happy. Bollen at el. (2011) found correlation between some of these states and Dow Jones Industrial Average (DJIA) with accuracy 87,6 % in predicting daily up and down changes. Alternatively, higher frequency of negatively ranked collective mood states, namely: anxiety, worry and fear leads according to Gilbert's & Harahalios's (2010) Fuzzy neural network model to downwarding pressure on S&P 500 Index.

Another effect caused by an existence of noise traders is under-reaction towards bad news and overreaction towards good news (Barberisa, 1998).

On the other hand we should take into account asymmetric fluctuation, when negative changes - crashes take short time, but the rise to the original price level takes significantly longer (Isen & Means, 1983).

According to Nofsinger (2015) stock market itself can be a direct predictor of social mood, despite the fact that earlier works on the same topic were not very useful in their prediction and only seldom had better predictive power than base model.

Porshnev et al. (2016) used ARMAX-GARCH model for returns modeling, in combination with lexicon-based approach with a list of emotional markers for sentiment analysis. This synthesis resulted in the most effective models in time series analysis, despite the fact it "allows to capture the main distinctive features of financial time series" (Horv & Kokoszka, 2003).

Li et al. implemented tensor-based learning technique based on fusion of sentiment analysis and opinions of professional obtained from financial news along with individual characteristics of companies (2016). This approach joining various sources of complex information is known as *mosaic* information space, where underlying interaction among sources can be found (Francis et al., 1997). Since textually dissimilar information may represent the same movement direction, super-feature vectors are created with the use of non-linear techniques as multi-layer neural network or kernel-based SVM in order to maximally preserve and strengthen interactions included in the space.

Beside other things, this analysis in its essence denies Malkiel's Random Walk hypothesis, which states that buy and hold strategy according to indices as S&P 500 is likely to outperform actively managed funds, whose managers are trying to pretend, that they are able to predict the future trend (1999). On the other hand, this theory was created at the time when no program trading existed and information was not spread so rapidly fast.

Another concept, we would like to disapproved is Efficient market hypothesis stating that the prices reflect all available information precisely at time news is made public, which implicates impossibility of buying stocks for value different from the fair one (Fama, 1965). Both mentioned concepts say that an attempt to predict market values to beat the market should fail, but later works proved them to be wrong. In the contrary to that, an experiment conducted by Lebaron revealed a lag in the time that information was introduced and the time the market would adjust itself to the new fair value (Sebastian, 2010).

3 Methodology and Evaluation

3.1 Research Questions and Hypothesis

RQ 1: Does the social media sentiment analysis have enough predictive power to estimate stock market trends for the next day?

RQ 1A: Can SA method be used as a predictor for the best time for selling stocks?

RQ 1B: Can SA method be used as a predictor for the best time for buying stocks?

RQ 2: Do individual characteristics of a company, like sector and key performance indicators (KPIs) influence the correlation?

3.2 Sentiment Classification

According to various surveys and companies (f. e. Biz360), rate of human concordance concerning sentiment of given text is between 70 - 79%, making a sentiment itself the most limiting factor. Concordance can be measured by weighted Cohen's κ, which accounts for degree of disagreements between annotators. Generally said, even in case of creating almost perfectly accurate model, approximately 3 people out of 10 might have absolutely different opinion about the sentiment of given information. Moreover, in case of financial messages sentiment quite often depends on previous expectations of investors. Exemplary case is message like "Year-to-Year profit increased by 10 %", which is generally positive on one hand, but when investors expected growth by 20 %, then this is in fact negative information and sentiment information would lead us to erroneous prediction. Despite these facts, we want to spread characteristics of "neutral words" with a purpose to eliminate miss-classification due to the noise and algorithm limitations. Sentiment analysis is mainly looking for the opinions, however every opinion needs to be linked with a target: for example the statement "I like Apple more than Samsung", is a positive message from Apple's point of view, but not from

Samsung's.

3.3 Role of Text Pre-Processing

Text pre-processing plays essential role in text mining analysis, especially in case of analysis of social media posts. The application of text pre-processing steps has one main reason and it is reduction of feature set, which might otherwise result in sparsity. Tweets are about everything and nothing at the same time, users create their own words, spelling shortcuts, tend to misspellings, slang words, insert URLs, genre specific terminology and abbreviations (Singh, 2016).

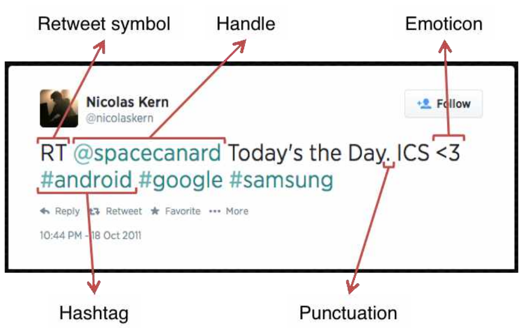

Figure 1: Illustration of a Tweet with various features

Source: *www.github.com/yogeshg/Twitter-Sentiment*

In the first step of text pre-processing, all letters are normalized to lower case, then each non-standard word needs to be treated specially before proceeding to analysis. The rule of thumb advises us to get rid of URL-links, diacritic symbols, punctuation, user references, duplicated Tweets and also stop-words. By stop-words are meant words like *and, the, of, is* etc., which are used for building of a sentence, but do not hold any sentiment information (Full list of excluded stop-words can be found in appendix A).

Next suitable step is to replace all negations (”not”, ”no”, ”never”,”without” etc.) by a single tag ”not”, preserved as Boolean variable in separate column for each Tweet, and consequently treat change in sentiment sign. We could not directly switch only sentiment sign, but also absolute value needs to be adjusted. For that purpose, we used lexicon with separately evaluated sentiment for sentences with and without negations. If the negation is directly included, we use its score, otherwise we only revert the sign of sentiment score, but it is needed to say that this case is relatively rare. We did not have to deal with it much.

In most cases, numbers are also removed, which seems to be inappropriate for posts with financial context, where numbers play crucial role as a part of information about performance efficiency of target company. We agreed to group numbers into special tag including also information, whether original number was positive or negative (Tellez et al., 2017).

Interesting feature, really useful for TSA is a use of emojis. Most of commonly used emoticons directly display frame of mind of its author. This can be easily used to assign the data-set of most frequently used emojis to their sentiment classification. This method proved to be holding significant explanatory power by various authors (Boia, 2012). We would like to mention and use for further work only 12 basic, most frequently used emoticons; half of them is used to express a positive mood: " :) ", " :-) ", " :)) ", " :D ", " =)) ", " :-))" and another half includes frowns like " :(", " :-(", " :/ ", ":-((", " :-P ", " :p " (Schnoebelen, 2012)

Table 1: Ranked Emoticons

Pos	Score	Neg	Score
:)	1,024	:(	-0,165
:-)	0,419	:/	-0,117
:))	1,519	:-(	-0,779
:-))	2,604	:-P	-1,254
:D	1,358	:p	-0,058
=))	2,056	:-((	-1,598

Another issue, needed to be treated, somehow, is an occurrence of multiple symbols, forming recognizable derived words. At this stage of the text pre-processing we get rid of symbols occurring in sequence of 3 characters and longer. This leads to a problem with transformation of specific words that can be interpreted two and more different ways. A typical example might be the word *GOOOOD*: it can mean either *Good* or *God*, a decision of which one was originally meant by an author is made based on occurrence of these words in training set. We agreed to remove only 4^{th} and each next repetition, and the most frequent words were subsequently added to our dictionary with tuned sentiment.

Another group of special words, abbreviations and slang words, counting expressions like "LOL" ("lots of laughs"), "AFK" ("away from keyboard") etc. needed to be examined individually. We agreed to annotate only the most frequent of them, and also add them subsequently to our evaluation corpus. Alternative possibility was using states from the internet sources like "The Online Slang Dictionary", but we decided not to go this way, because majority of terms found there were irrelevant for our purpose.

Up to now, we discussed general issues insensitive to the language used, but at this moment, we need to restrict our analysis only to English written text. Thanks to an increasing quality of online text translators, we might be able to also translate non-English texts into English and then proceed with analysis also for other languages. However this process is not the most suitable one, with adjusted weights of non-English text, we might use also information obtained from such texts.

Different commonly used pre-processing features are stemming, punctuation and "italic type" intensifiers, but we decided not to cover those in this thesis for simplification. Main purpose of stemming is removal of a major type affixes and/or prefixes, followed frequently by process of normalization. The first stemming algorithm was coded by J. B. Lovin in 1968, but currently more frequently used is Porter's Stemmer that is more accurate and also faster (Garg, 2014). Punctuation intensifiers contain for example characters like "!!!" resulting in intensification of sentiment of given sentence, but since we are working on word level, it does not make sense to apply this in our work. The same is the case of "italics type" intensifiers, where for example highlighted sentiment "SO GOOD" should have more positive score than casual "so good" etc.

We also used another approach: Inverse document frequency (idf), which is part of tf-idf statistics and is used to measure importance of a word for a document in a collection. *"The idea of tf-idf is to find the important words for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of document."* (Silge & Robinson, 2017), which leads to elimination of the most commonly used words. For Tweets Idf is counted using the following equation:

$$idf(Word) = ln\left(\frac{n_{\text{ Tweets}}}{n_{\text{ Tweets containing the Word}}}\right) \tag{1}$$

Firstly all Tweets are evaluated using slightly adjusted financial Lexicon and than tf-idf approach is used to identification of potentially useful terms, which are joined with original set.

3.3.1 Steps of Text Pre-Processing

- Remove all duplicated Tweets and substitution of this information by number of retweets instead of duplicated text itself.
- Evaluate sentiment of emoticons, it there are any included. After evaluation, emoticons are deleted from the corpus.
- Normalize to lowercase.
- Get rid of url-links, diacritics symbols, user-reference, punctuation and stop-words.
- Tokenize all negation terms by "not", values by "+/- val."
- Normalize character sequences - Delete sequences of repetitive characters with length higher than 3. Deciding whether it makes a sense to treat such words by boosting their sentiment or not according to significance of proportional representation.

3.4 Machine Learning & Classifiers

After the pre-processing is completed, it is necessary to represent features remaining in the corpus. We decided to use n-gram representation. All possible sequences of n-words are taken from each Tweet and compared with corpus created during the training of the model afterwards.

Letś take a following text as an example:

JimCramer · @*jimcramer* · Jul 10
$AAPL rallying despite the worst press imaginable.

After the pre-processing is done, we are left with the message:

"rallying despite worst press imaginable"

Being aware of the tri-gram word features we obtain the following overlapping triplets: "rallying despite worst", "despite worst press", "worst press imaginable". Some methods are also starting with getting rid of another parts of speech, but we suppose that for financial Tweets adjectives are sometimes even more important than nouns. In our analysis, we agreed to take advantage of using special financial lexicon containing not only bi-grams, but uni-grams as well. Bi-grams have better ability to capture relationships between the words that are frequently used in sequence or co-occurrence within the sentence rather than individual words and also depict differences in word-level emotion of the word. Moreover, number of possible combinations is not so high in comparison with higher n in case of n-grams.

On the other hand, much wider range of possible combinations exists and the whole processing of more extensive lexicon takes significantly longer, which might be disadvantage for intra-day predictions and large set of Tweets. Uni-grams and bi-grams contained in corpus of Tweets are compared with the lexicon of evaluated financial uni-grams and bi-grams, where the first iteration is looking for bi-grams only and then, if bi-gram is found, the involved words are excluded from the second iteration looking for corresponding uni-grams.

We agreed to use, as a baseline, stock market lexicon created by Oliveira et. al (2016) using large data-set obtained from micro-blogging services, especially from StockTwits. This lexicon contains not only affirmative sentiment scores for each n-gram, but also score for negated context. Lexicon contains 20 550 expressions ranked in interval [-10,10] and is well balanced with 10 546 positive and 10 005 negative expressions included . Whereas number of items in each category is balanced, the same does not hold about mean sentiment score within both categories. We agreed to move our sentiment score by the obtained mean score, subsequently in order to have mean value inside neutral sentiment range. Expressions found in Tweets, but not included in our lexicons are scored with 0.

This method is far from perfect, but we hope that approach will be sufficient enough for our purpose. We still need to question stability of our domain specific lexicon. Similar steps were executed by Flekova et al. (n/a) with accuracy around 0,7, which can be counted as success. Important piece of information to be taken from their paper is a conclusion that introduction of new errors by usage of bi-grams is caused by remaining ambiguity of more complex linguistic units.

The most common data structure used for R-Text mining is document-term matrix (DTM). For these matrices, one document corresponds to one row, the same holds about columns and terms. Count of occurrences within given text can be found in values section of the matrix. In our case, we prepared DTM where each row corresponds to trading day and columns contain subset of uni-grams and bi-grams from financial lexicon AZFinText (Schumaker et al., 2012). Our DTM matrix needed to be transferred to *Tidy* format so we were able to use R-package *Tidytextmining*. Part of NLP study inquires into methods based on word-co-occurrence. For this case, we decided to use Kilgarriff et al.'s (2004) Lexicographer's Mutual Information (LMI) to evaluate bi-grams which frequently co-occure with positive/negative mood. This approach is quite handy, because it does not discriminate frequently used combinations of words. It can be defined as:

$$LMI(w,c) = log_2\left(\frac{f(w,c)}{f(w) * f(c)}\right) * f(w,c) \tag{2}$$

where f() denotes frequency of word (w), context (c) and both of them together (w,c). LMI is computed over corpus divided by positive and negative bi-gram score. Prior the computation we also got rid of noise by removing duplicity and subsequently we computed semantic orientation of each bi-gram: $LMI_{SO} = LMI_{pos} - LMI_{neg}$. The subject of our interest are bi-grams evaluated as positive or negative, which have higher score of an opposite direction on word-level. The most frequent results were then added to our lexicon of financial bi-grams with a corrected score. Bi-grams that would be incorrectly classified are for example: *cold person/beer, sincere condolences, cloud computing, guilty pleasure/feeling, dark chocolate/thoughts etc.* Without this separated treatment such bi-grams might harm our sentiment predictions.

3.5 Tests

3.5.1 Granger Causality

Despite the fact that most of the previous papers found correlation between TSA and Stock market returns, we need to make sure, that we are using the relationship between variables in the correct direction and that lagged terms have explanatory power for our dependent variable. To predict future values y_t, controlling for past y and past z, we test for Granger causality. Variables y and z are part of following general equation:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + \alpha_2 y_{t-2} + \gamma_2 z_{t-2} + \cdots \tag{3}$$

Generally, z Granger causes y if:

$$E(y_t \mid I_{t-1}) \neq E(y_t \mid J_{t-1}) \tag{4}$$

where J_{t-1} contains information on past y only, whereas I_{t-1} also includes information on z. If this equation holds, past z has an effect on predicting y_t along with addition of past y. For our model, we need to test hypothesis that y does not Granger cause z, more precisely that any lag of z should have zero population coefficients (Wooldridge, 2013). Join significance of lag-terms can be tested, for example, by an F-test in casual form or robust form in case of heteroskedasticity.

In order to avoid inclusion of the reverse causation, we use approach introduced by Brown & Cliff (2004). We only kept those emotional markers with β significant and $\tilde{\beta}$ insignificant at the 5% level according to F-test. Only markers and lags, which pass the test, can be used as explanatory variables in the model.

3.5.2 Diebold-Mariano test

Diebold-Mariano test (DM) for predictive accuracy is a useful tool for comparison of two or more time series models available for prediction of variable of interest. In conservative way it can be applied well to short-horizon forecasts, making it exactly eligible for our case (Giacomini & White, 2003). DM forecasts errors ($e_{it} = \widehat{y_{it}} - y_t$) and make assumptions based directly on those forecast errors. It requires the loss differential to be co-variance stationary. A symmetric squared-loss function g(•) is created from forecast errors. DM assumptions are following:

$$\text{Assumptions DM:} \begin{cases} \forall \text{t: } E(d_{12,t}) = \mu \\ \forall \text{t: } cov(d_{12,t}, d_{12,(t-\tau)}) = \gamma(\tau) \\ 0 < var(d_{12,t}) = \sigma^2 < \infty \end{cases}$$

In fact, DM is simply an asymptotic z-test of the hypothesis regarding mean of loss differential to be zero for observed time series. Some complications might occur due to the serial correlation between forecast errors and loss differentials. One alternative for loss function is figure $g(e_{it}) =$

$exp(\lambda e_{it}) - 1 - \lambda e_{it}$, where λ is a positive constant. Our null hypothesis is defined as: $H_0 : E(d_t) = 0 \; \forall t$ (i.e. both models have the same accuracy), where d_t is difference of g-functions constructed for models under the comparison (Diebold, 2013). Description of all necessary steps of computation of this test is out of the scope of this thesis. Since we are interested only in forecasts of 1 step ahead (h=1). DM can be then written as follows:

$$DM = \frac{\overline{d}}{\sqrt{\frac{2\pi \widehat{f}_d(0)}{T}}} \tag{5}$$

where $\widehat{f}_d(0)$ is a consistent estimate of $f_d(0)$, which is population mean and $\widehat{f}_d$ is the sample mean of the loss differential. The last term T is sample size of t.

4 Data

4.1 Twitter

We filtered only English Tweets for simplification, so they can be used directly for the purpose of TSA using lexicon-based algorithm. Otherwise we would be forced to deal with each language separately or translate each non-English Tweet , for example via Google translator (Sample data can be found in Appendix B). Although its quality increased rapidly in recent months thanks to the implementation of updated learning algorithm, the results might still be influenced by not exactly precise translation. The choice of English Tweets only should not cause any troubles, because we are comparing sentiment of Tweets with an American stock market index S&P 500. Also it could be supposed that most of business oriented texts are written in English.

It seems reasonable to suppose that the mood of the tweets reflects general feelings of society toward selected topic. For extraction of required Tweets connected to the selected companies, we used a feature recently added to Twitter called cash-tag. Cash-tag is a notation used for tweets linked to the stock market. It is constituted by prefix $ and stock quote (f. e. $AAPL, $GOOG, $GSPC etc.).

Due to the recent Twitter policy changes, data sets containing Tweets should not be publicly available, which significantly complicates our analysis. To obtain data-set applicable for model training, we needed to manually annotate at least fraction of data. That basically means that we want to test lexicon-based algorithm against set corresponding to our perception of sentiment, which might differ from professional perspective though. These static data-sets were in the past often used only for training and testing the model using supervised techniques.

Data for predictive analysis are acquired directly from Twitter.com via Streaming API, which allows to write a code to download Tweets meeting the requirements directly into R Studio and to immediately process the data. Twitter policies placed currently relatively strict limitation on number of

Tweet and period of last few days when Tweets are available, regrettably this period is too short to be used for our purpose.

After the TSA of each individual Tweet was executed, we created time-series for each stock title as well as for market index, with calculated weighted averaged sentiment score for each day. Tweets with most re-tweets will be advantaged, because they should depict the collective sentiment more precisely. The weighted sentiment is calculated as follows:

$$TS_t^{\pm} = \frac{\sum_{0 \leq i \leq n} SENT_{i,t} * (RETW_i + 1)}{\sum_{0 \leq i \leq n} (RETW_i + 1)} \tag{6}$$

where n denotes set of Tweets about selected item at time t and i is single item from set n. Variable $RETW_i$ is approximation of popularity and influentialness of given Tweet, for simplification we omitted time component t of this variable, because we do not want to keep duplicity in our set and it is also rational to suppose that most of the re-tweets comes in the day of publishing. Constant *+1* is added to re-tweets in order to count also the original Tweet.

Despite the fact that streaming API in R is only able to download Tweets for the last few days, we downloaded data-sets containing financial Tweets about 102 American companies covering period from *02.04.2016* to *15.06.2016*, which we hope will be sufficient amount for training of model. From already limited data set, we found amount of records insufficient for majority of companies. Data-sets with less than 10 000 Tweets for given period were discarded. Unfortunately, by this restriction we lost 88 of data-sets and only 14 (13) of them remained. At first glance, there are mainly huge and stable companies, which shares might not have a tendency to fluctuate due to sentiment changes.

Data obtained from Twitter needed to be joined with Stock market information, which requires the dates to correspond. For NYSE core trading session starts at 9:30 a.m. and ends at 4:00 p.m. Eastern time (ET), but Tweets can be published during the whole day. Firstly we agreed to unify time-zones of all Tweets to ET and consequently shift the date for Tweets that had no chance to influence closing price of given day. We suppose that some of the crucial news might be announced after trading day is over, resulting in miss-classification.

Table 2: Tweet datasets with sufficient number of Tweets

Stock Quote	**Company Name**	**Tweet Count**
$AAPL	Apple, Inc.	166 631
$AMZN	Amazon.com, Inc.	57 378
$CSCO	Cisco Systems, Inc.	13 283
$FB	Facebook, Inc.	93 898
$GILD	Gilead Sciences, Inc.	16 146
$GOOG	Alphabet, Inc.	37 642
$GOOGL	Alphabet, Inc. *(dup.)*	29 385
$MSFT	Microsoft Corporation	43 060
$NFLX	Netflix, Inc.	37 083
$NVDA	NVIDIA Corporation	10 933
$PCLN	The Priceline Group, Inc.	11 896
$SBUX	Starbucks Corporation	13 941
$TSLA	Tesla, Inc.	82 477
$YHOO	Yahoo! Inc.	17 614

4.2 Stock Market

The second time-series covers stock market data for selected companies in a given period. The data is directly downloaded with the help of R-Package *quantmod* from Yahoo! Finance. The same method is also used for stocks listed in stock market index S&P 500-every time it is sufficient to insert Yahoo! ticker and data are directly downloaded within few seconds.

Metric used for computation of stock returns during the training of the model is the natural logarithm of the first difference of the adjusted closing stock prices, which can be easily used even for dividend stocks. We stick to daily returns only and thanks to that matter, our model might be performing better. Log-returns can hardly be combined with fat-tailed distributional assumptions even for short-term data, which we do not have ambitions to cover. Returns are calculated by following formula:

$$r_t = log\left(\frac{p_t}{p_{t-1}}\right) \tag{7}$$

Missing data for non-trading days were approximated using a simple technique described by Goel (2011). The technique is based on observations of shape of stock price function, which most of the time follows a continuous concave function. The gap is filled from the middle out, where the first approximated value is mean of the endpoints.

4.2.1 Stock Market - Selection

Originally, we hoped to start our analysis with full set of 500 companies included in index S&P 500, but then limited the selection only to companies, which do not pay out dividends. We decided to do so because changes in price are mainly not connected with sentiment and we would be forced to treat them manually in case we did not exclude them from selection. At the end, we ended up just with handful of stocks, which hopefully have certain characteristics that might indicate sentiment-sensibility for example lack of data for fundamental analysis, high volatility and relatively high price in comparison with negative profit. Here we might deal with the possible obstacle that, according to authors as Zhang et al. (2016), observation of individual stock does not contain enough information for prediction of future behaviour. Despite this idea might lead to remarkable results, we were not able to obtain sufficient data to proceed in this way, thus we decided to analyze available data and make conclusion based on their analysis only.

5 Empirical Model

5.1 General Model

We propose the following model capturing the overall dynamics with 5 lag-terms, considered by most of the current papers to be the most significant.

$$r_t = \beta_0 + \beta_1 r_{t-1} + \gamma_1 TS^{\pm}_{t-1} + \cdots + \gamma_5 TS^{\pm}_{t-5} + \beta_2 S\&P500_{t-1} + \mu_t \qquad (8)$$

Where r_t is logarithmic form of returns at time t, $TS^{\pm}_t$ is a sentiment score with values, optimally, in interval $[-1, 1]$ at time t, $S\&P500_{t-1}$ covers logarithmic change in the value of market index at the time *t-1* where all selected companies belong to. This term should serve us like a proxy for general sentiment and overall condition of economy. At the end of equation there is an error term μ_t referring to unexplained variance in returns. This model assumes symmetrical relationship for negative and positive sentiment, which might not be the most appropriate way. In case of lag, the model is working on daily basis, t is measured in calendar days. For intra-day analysis, number of lag-parameters would need to be significantly higher.

As a next step, we agreed to place a threshold on $TS^{\pm}$, for the reason that we suppose the sentiment does not influence stock price in any way during the calm times. If the threshold is not surpassed, the model supposes constant growth pace or stagnation, depending on the performance during the previous trading day. Model with all lag-terms equal to zero is considered to be our benchmark model. This should also leads to the reduction of noise. We define our dummy threshold term Q as following:

$$Q_{t-i} = \begin{cases} 1 & \text{if } C_{t-i} > 1{,}2 \times \text{10-days simple moving average} \\ 0 & \text{if } C_{t-i} \leq 1{,}2 \times \text{10-days simple moving average} \end{cases}$$

Our original intention was to use 30-days moving average, but we experienced trouble with obtaining data-set covering long-enough period to implement period longer than 10 days, since the Tweets from the beginning of the period need to be discarded due to the unavailability of moving average value. Model with active threshold is defined as follows:

$$r_t = \beta_0 + \beta_1 r_{t-1} + \gamma_1 Q_{t-1} TS^{\pm}_{t-1} + \cdots + \gamma_5 Q_{t-5} TS^{\pm}_{t-5} + \beta_2 S\&P500_{t-1} + \mu_t \quad (9)$$

5.2 Sign Sensitive Model

It is rational to suppose, that stock market participants do not react in the same way on positive and negative news. Most of the time, serious negative news results in a rapid drop, whereas positive news does not immediately lead to such a rapid growth. As a results of this - γ coefficients might have different absolute values. For that purpose, we divided our sentiment data into two groups: TS^{+} and TS^{-} with the same threshold as original model. The resulting models are defined as follows:

$$r_t = \beta_0 + \beta_1 r_{t-1} + \gamma_1 Q_{t-1} TS^{+}_{t-1} + \cdots + \gamma_5 Q_{t-5} TS^{+}_{t-5} + \beta_2 S\&P500_{t-1} + \mu_t \quad (10)$$

$$r_t = \beta_0 + \beta_1 r_{t-1} + \gamma_1 Q_{t-1} TS^{-}_{t-1} + \cdots + \gamma_5 Q_{t-5} TS^{-}_{t-5} + \beta_2 S\&P500_{t-1} + \mu_t \quad (11)$$

the last alternative, which might be used, is to join non-corresponding parts of the last two equations together. That would result in inclusion of TS^{+} and TS^{-} next to each other in the model, not only getting single sentiment score for each lag. This comes from an idea, that coefficients for negative sentiment at time t are not the same as coefficients for positive sentiment, due to the standard shape of individual utility function - absolute value of the change of the utility is not the same if one loses $ 1 and if he gains $1.

5.3 Regression

Models are estimated by Ordinary Least-Squares Regression (OLS), which is fully sufficient for this type of models. Also its results are easily interpretable and testable. OLS technique for linear modelling can be applicable on multiple continuous explanatory variables (X) with a single response variable (Y). It can be represented mathematically as: $Y = \alpha + \beta X$, where α is an intercept and β is a vector of individual slopes of best-fit lines for each explanatory variable. For testing part of our modeling we compared predicted $\tilde{Y}$ with observed Y on sample data, this is known as deviation of Y. These deviations are squared and their sum, known as the residual sum of squares (RSS) and provides information of model-fit for the regression. The deviance is important measure of predictive power of an additional variable, when notable reduction of deviance indicates significant effect on prediction of Y.

We now need to verify assumptions of OLS model according to Wooldridge (2013). We hope the model is linear in all parameters, but we already know that our sample is not random, because our data belongs to time-series group. Also variation is obviously included variation in our explanatory variable and our error term u has an expected value of zero given any explanatory variable.

6 Empirical Results

Empirical results are highly influenced by unavailability of sufficient amount of data. Initially, we wanted to train and test model for each company with at least 10 000 Tweets during the given period with a majority of days with positive returns. For some companies we obtained only negative or positive sentiment scores above threshold, thus it does not make a sense to base predictions on coefficient trained on such data. It is quite certain that this trend cannot last for ever. Unfortunately after first few models were trained, it became obvious that it is not possible to use them for making any prediction, since they are not working well even on training data, mainly due to various sentiment scores included in the lexicon.

6.1 Apple, Inc.

First of all, we evaluated sentiment and ran OLS regression on our widest data-set, the one containing Tweets about \$AAPL, without distinction between positive and negative we obtained following coefficients:

$$\begin{aligned} r_t = -0,017 + 0,441 r_{t-1} + 0,591 Q_{t-1} TS^{\pm}_{t-1} - 0,563 Q_{t-2} TS^{\pm}_{t-2} \\ + 0,991 Q_{t-3} TS^{\pm}_{t-3} - 1,239 Q_{t-4} TS^{\pm}_{t-4} - 1,096 Q_{t-5} TS^{\pm}_{t-5} \\ - 0,477 S\&P500_{t-1} \end{aligned}$$

with only r_{t-1} statistically significant at 5% level. Only two another variables are relatively close to the significance level: $TS^{\pm}_{t-3}$ and $TS^{\pm}_{t-4}$. The most curious thing is overall positive sentiment despite the fact that majority of tracked period is connected with decrease of value of \$AAPL shares. The same holds even if the threshold is omitted. What corresponds to our initial theory is sign of most of sentiment lag-terms. Opposite sign might indicate return to fair-value, after over-reaction period is over, but this is just our personal estimate without any statistical support behind.

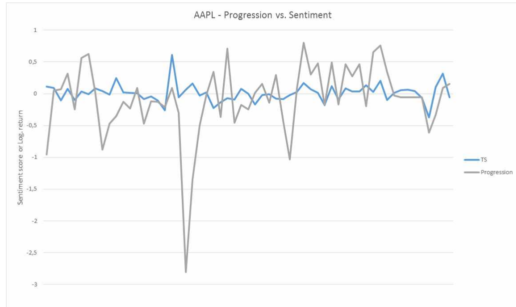

Figure 2: AAPL - Progression vs. Sentiment score

Source: *Author's computation*

What was not obvious from estimated coefficients might be quite clearly seen in case of graphical representation of progression of closing stock price next to Twitter sentiment score, where both of the curves quite often follows the same shape and direction. What seems to be probable is finding that positive sentiment only minimizes losses, not providing gains and vice versa.

Despite the fact that no improvement was achieved by limitation to positive sentiment only, we found significant improvement by limitation to negative sentiment only in combination with omitting variable recording progression of index S&P 500 during the previous day. Without this variable, first-lagged term was significant on 5% level as well as performance during the previous day. Another unsuccessful attempt was connected with lowering of threshold to 10 % above 10-days moving average.

6.2 Amazon.com, Inc.

Secondly, we ran regression on data for $AMZN, where we again found, in case of general model, two statistically significant variables: $TS^{\pm}_{t-4}$, with reverse sign than dependent variable and $S\&P500_{t-1}$, with coefficient equal to 1,1, which might indicate behaviour similar to the whole index, with little overreaction. Despite insignificance of remaining lag-terms, it is interesting to see change in sign between second and third day, which might be expected behaviour as a result of correction of high optimism. We guess that problem here is connected with the approach used for sentiment analysis, where for days when Tweets contains relatively high amount of terms included in the lexicon, gained disproportionately high or low sentiment score, which is in combination with limited amount of data quite serious problem.

Overall performance decreased by limitation to positive sentiment only, but significantly increased with the use of only negative sentiment. If we took into account only negative scores, we found four variables significant at 5 % level: Lagged sentiment at time *t-1*, *t-2*, *t-4* and also $S\&P500_{t-1}$. This outcome might be surprising, because in given period, we obtained negative score along with surpassed threshold for 3 days only. It would be incorrect to neither make any predictions based on this model, nor it makes a sense to combine negative and positive models, when none of them contain a valuable information.

6.3 Facebook, Inc.

According to our expectation, we obtained the best results with general model in this case. After few another adjustments, we found the best performance for case when returns during the last day were excluded from the model and threshold set to 10 %. In this case, all variables, except from the fourth lagged term are statistically significant at 5 % level. We obtained following coefficients:

$$
\begin{aligned}
r_t = &-0,005 - 0,052 Q_{t-1} TS^{\pm}_{t-1} + 0,083 Q_{t-2} TS^{\pm}_{t-2} + 0,091 Q_{t-3} TS^{\pm}_{t-3} \\
&- 0,020 Q_{t-4} TS^{\pm}_{t-4} - 0,075 Q_{t-5} TS^{\pm}_{t-5} - 0,477 S\&P500_{t-1}
\end{aligned}
$$

Here it does not make sense to distinguish between positive and negative $TS^{\pm}$, because we obtained positive score in everyday of the tracked period. What is interesting here is the fact that all coefficients of lagged sentiments are relatively close to each other, which might be caused by not so high variance in individual sentiment scores, compared to previous two cases.

6.4 Tesla, Inc.

The last tested company was $TSLA, where no evidence of statistical influence of sentiment on stock returns was found. The only significant variable, at 5 % level was performance during previous day, which indicated more or less constant drop in price during observed period.

6.5 Tests

Since no model for any data-set resulted in statistically significant variables, we suppose that it does not make much sense to run Granger Causality test, because even if it would end with positive results, it seems that our models have hardly any predictive power. On the other hand, DM test is powerful tool for features selection, however in case of lack of data it might even worsen predictive power of the model. We used pseudo-DM test and tried in each case to omit either lagged term tracking performance during the previous period or trend of stock market index during the previous period and sometimes even both without any significant improvement in model performance. Sentiment variables and their significance seems to be driven by coincidence, because their (in)significance seems to be more or less random and it would be mistake to omit lag-terms inside the row as a result of testing on insufficient data.

6.6 Evaluation of RQs

Regrettably, due to our set up and lack of data, we were not able to meaningfully evaluate not even a single research question we laid in the beginning of this thesis, despite we observed successful answers for our basic research questions obtained by other researchers. We still need to conclude that in our analysis, social media sentiment does not have enough predictive power to estimate the stock market trends for the next day, regardless of individual characteristics of a company. Due to the conclusions mentioned before, we decided not to base any trading strategy on our findings due to their unsatisfactory performance.

Conclusion

We conducted not very successful analysis of Twitter sentiment and its relation to stock market. Our models consisted of 5 lagged sentiment terms, lagged return of given stock and corresponding index during previous day. Sentiment score was obtained for each Tweet using lexicon-based method, with domain-specific lexicon. After we obtained the score, we weighted it for each Tweet according to its number of re-tweets compared to rhe total number of Tweets, which tends to produce scores with high volatility. This in combination with low amount of data led to inconsistent and statistically insignificant results. We tried to normalize sentiment scores without any success. It would not be appropriate to make any predictions based on our model with coefficient trained on data-sets we used. Perhaps only surprising thing is similarity of graphical representation of sentiment scores and returns, which might indicate, that overall effect of sentiment is reflected into stock price within trading day already. We underestimated given topic, where lexicon-based analysis seems to be inappropriate, mainly due to wide range of possible scores and high fluctuation of sentiment without influence on the stock market side.

Even if we were able to create faultless sentiment evaluation incorporated in predictive model, the difficulty is in inability to capitalize on the behaviour of individual traders, whose behavioural patterns change over time, and also on unpredictable external influences.

In the future work we might try to use some deep-learning algorithms and also optimize our current algorithm and lexicon to be able to proceed predictions on intra-day basis and use wider range of sources (f.e. Google trends, Instagram and news headings) along with the TSA. We need to get rid of limitation caused by insufficient knowledge of advanced algorithms and other programming languages as Python and MatLab, which are in some cases more powerful, faster and provide ready to use packages and also allow to download data for longer periods. We had various different ideas, that might provide interesting results, but we were not able to transfer them successfully into the code and subsequently process them in analysis due to these limitations. It might be also interesting to build a model evaluating sentiment sensitivity according to various metrics and features of company characteristics. Nevertheless, even if someone managed to create a powerful predicting model, able to minimize risk and maximize profits, such research might never get published. We suppose that, most of the current researches might just become a revers case of file drawer effect.

References

[1] BARBERISA, N. (1998): ”A model of investor sentiment”. Elsevier

[2] BENGIO, Y. (2013): ”Deep Learning of Representations: Looking Forward.” Universite de Montreal. Canada

[3] BOIA, M. (2012): ”Context Sensitive Sentiment Analysis.”

[4] BOLLEN, J., H. MAO & X. ZENG (2011): ”Twitter mood predicts the stock market.” Journal of Computational Science. 2(1). 1-8

[5] BROWN, G. W. & M. T. CLIFF (2004). ”Investor sentiment and the near-term stock market.” Journal of Empirical Finance, 11(1): 1-27

[6] CHOI, J. & C. CARDIE (2009): ”Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification.” In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP, pages 590- 598

[7] COOK, P. & S. STEVENSON (2010): ”Automatically identifying changes in the semantic Orientation of words.” In Proceedings of the Seventh International Conference on Language Resources and Evaluation. pages 129-149

[8] DIEBOLD, F. X. (2013): ”A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests.” Journal of Business and Economic Statistics

[9] EIJI, A., S. MASKAWA & M. MORITA (2011): ”Twitter catches the flu: detecting influenza epidemics using Twitter.” Proceedings of the conference on empirical methods in NLP. Association for Computational Linguistics

[10] FAMA, E. (1965): ”The behavior of stock-market prices.” Journal of business. 38(1):34

[11] FLEKOVA, L., E. RUPPER & D. PREOTIUC-PIETRO (*n/a*): ”Analysing domain suitability of a sentiment lexiconb yidentifying distributionally bipolar words.”

[12] FORTUNY, E. J., T. SMEDT, D. MARTENS & W. DAELEMANS (2013): "Evaluating and understanding text-based stock price prediction models." Elsevier

[13] FRANCIS, J., H. J. DOUGLAS & D. PHILBRICK (1997): "Management communications with securities analysts." J. Account Econ. 24(3): 363-394

[14] GARG, Y. (2014): "Sentiment analysis of Twitter feeds." Indian institute of technology. Delhi

[15] GIACOMINI, R. & H. WHITE (2003): "Tests of Conditional Predictive Ability." Working Paper No. 572, Boston College

[16] GIDOFAVLI, G.(2001): "Using News Articles to Predict Stock Price Movements." Department of Computer Science and Engineering, University of California, San Diego

[17] GILBERT, E. & K. KARAHALIOS (2010): "Widespread Worry and the Stock Market." In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp.58-65

[18] GO, A., R. BHAYANI & L. HUANG (2009): "Twitter sentiment classification using distatnt supervision." vol. 150. Ainsworth Press

[19] GONZALEZ-IBANEZ, R., S. MURESAN & N. WACHOLDER (2011): "Identifying sarcasm in Twitter: a closer look." ACL-HLT 11, pages 581-586, Portland, Oregon, USA

[20] HORV, L. & P. KOKOSZKA (2003): "Garch processes: structure and estimation." Bernoulli. 9 (2). 201/227

[21] IKEDA, S., H. TAKAMURA, L. RATINOV & M. OKUMURA (2008): "Learning to shift the polarity of words for sentiment classification." In Proceedings of the International Joint Conference on Natural Language Processing, IJCNLP, pages 296-303

[22] ISEN, A. M. & B. Means (1983): "The influence of positive affect on decision-making strategy." Social cognition, 2 (1)

[23] JELVEH, Z. (2006): "How a Computer Knows What Many Managers Don't." The New York Times

[24] KADILI, A. (2015): "Predictability of stock returns of financial companies and the role ofinvestor sentiment: A multi-country analysis." Journal of Financial Stability

[25] KEARNEY, C. & S. LIU (2014): "Textual sentiment in

nance: A survey of methods and models." International Review of Financial Analysis. 33

[26] KILGARRIFF, A., P. RYCHLY, P. SMRZ & D. TUGWELL (2004): "Itri-04-08 the sketch engine." Information Technology. 105

[27] LEPORI, G. M. (2015): "Positive mood and investment decisions: evidence from comedy movie attendance in the us." Research in International Business and Finance. 34

[28] LI, Q., J. WANG, F. WANG, P. LI , L. LIU & Y. CHEN (2017): "The role of social sentiment in stock markets: a view from joint effects of multiple information sources." Springer

[29] LIEBRECHT, C., F. KUNNEMAN & A. V. BOSCH (2013): "The perfect solution for detecting sarcasm in tweets #not." In Atlanta, Georgia, USA

[30] LOEWENSTEIN, G. (2010): "Emotions in Economic Theory and Economic Behavior." The American Economic Review, 90 (2), 426

[31] MAGDY, W., H. SAJJAD, T. EL-GANAINY & F. SEBASTIANI (2015): "Distant Supervision for Tweet Classification Using YouTube Labels."

[32] MAGDY, W. & T. ELSAYED (2014): "Adaptive method for following dynamic topics on Twitter." In ICWSM

[33] MALKIEL, B. G. (1999): ”A random walk down Wall Street: including a life-cycle guide to personal investing.” New York: Norton. c1999. ISBN 0-393-04781-4.

[34] MOHAMMAD, S., S. KIRITCHENKO, & X. ZHU (2013): ”Building the state-of-the-art in sentiment analysis of tweets.” NRC-Canada, Atlanta, Georgia, USA

[35] MITRA, R., B. RIEDL, C. BIEMANN, A. MURKHERJEE & P. GOYAL (2014): ”That's sick dude!: Automatic identification of word sense change across different timescales.” pages 1020-1029

[36] MOILANEN, K. & S. PULMAN (2008): ”The good, the bad,and the unknown: Morpho syllabic sentiment tagging of unseen words.” In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, ACL, pages 109-112

[37] NAKOV, P., S. ROSENTHAL, Z. KOZAREVA, V. STOYANOV, A. RITTER & T. WILSON (2013): ”SemEval-2013 Task2: Sentiment Analysis in Twitter”

[38] NI, Z. (2015): ”Investor sentiment and its nonlinear effect on stock returns-New evidence from the Chinese stock market based on panel quantile regression model.” Elsevier

[39] NIELSEN, F. (2011): ”AFFIN.” Informatics and Mathematical Modelling, Technical University of Denmark

[40] NOFSINGER, J. R. (2015): ”Social Mood and Financial Economics.” Journal of Behavioral Finance, 6 (3)

[41] OLIVEIRA, N., P. CORTEZ & N. AREAL (2016): ”Stock market sentiment lexicon acquisition using microblogging data and statistical measures.” Decision Support Systems 85: 62-73

[42] PAK, A & P. PAROUBEK (2010): ”Twitter as a corpus for sentiment analysis and opinion mining.” Conference paper

[43] PARTELOVÁ, M. (2017): "Presidential rhetoric, sentiment and their relation to stock markets." Bacherols Thesis. Prague

[44] PALLIER, G., R. WILKINSON, V. DANTHIIR, S. KLEITMAN, G. KNEZEVIC, L. STANKOV & R. D. ROBERTS (2012): "The role of individual differences in the accuracy of con

dence judgments." The Journal of General Psychology

[45] PANG, B., L. LEE & S. VAITHYANATH (2012): "Thumbs up? Sentiment Classification using Machine Learning Techniques." Available from www:`http://www.cs.cornell.edu/home/llee/papers/sentiment.home.html`

[46] PORSHNEV, A. V., V. L. LAKSHINA & I. E. RED'KIN (2016): "Could emotional markers in Twitter posts add information to the stock market ARMAX-GARCH model." National research university higher school of economics

[47] REN , Y., Y. ZHANG, M. ZHANG & D. JI (2016): "Context-sensitive twitter sentiment classification using neural network." Thirtieth AAAI Conference on Artifical Intelligence

[48] SCHUMAKER, R. P., Y. ZHANG, C. N. Huang & Chen, H. (2012). "Evaluating sentiment in financial news articles." Decision Support Systems. 53(3). 458-464

[49] SCHWARTY, H. A., J. EICHSTAEDT, L. DZIURZYNSKI, E. BLANCO, M. L. KERN, S. RAMONES, M. SELIGMAN & L. UNGAR (2013): "Choosing the right words: Characterizing and reducing error of the word count approach." In Proceedings of the Second Joint Conference on Lexical and Computational Semantics, SEM, pages 296-305

[50] SEBASTIAN, M. (2010): "Modelling the Stock Market using Twitter." University of Edinburgh

[51] SCHNOEBELEN, T. (2012): ”Do you smile with your nose? Stylistic variation in Twitter emoticons.” University of Pennsylvania Working Papers in Linguistics

[52] SILGE, J. & D. ROBINSON (2017): ”Text Mining with R: A Tidy Approach.”

[53] SINGH, T. & M. KUMARI (2016): ”Role of Text Pre-Processing in Twitter Sentiment Analysis.” National Institute of Technology

[54] TELLEZ, E., S. MIRANDA-JIMENEZ, M. GRAFF, D. MOCTEZUMA, R. R. SUAREZ & O. S. SIORDIA (2017): ”A simple approach to multilingual polarity classification in Twitter.” Elsevier. Available from www: `http://www.sciencedirect.com/science/article/pii/S0167865517301721`

[55] TU,W., D. CHEUNG & N. MAMOULIS (2015): ”Time-sensitive opinion mining for prediction.” In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

[56] DE VRIES, L., S. GENSLER, P. S. H. LEEFLANG (2012): ”Popularity of brand posts on Brand Fan pages: an investigation of the effects of social media marketing.” J. Interact. Mark

[57] WOOLDRIDGE, J. M. (2013): ”Introductory Econometrics: A Modern Approach.” Fifth Edition

[58] ZHANG, G., L. XU & Y. XUE (2016): ”Model and forecast stock market behavior integrating investor sentiment analysis and transaction data.”

[59] Investopedia (2017), available from `www.investopedia.com/terms/e/efficientmarkethypothesis.asp`

List of tables and figures

Appendix A

List of stop-words

a	believe	ever	indicates	nobody	saw	then	viz
able	below	every	inner	non	say	thence	vs
about	beside	everybody	insofar	none	saying	there	want
above	besides	everyone	instead	noone	says	there’s	wants
according	best	everything	into	normally	second	thereafter	was
accordingly	better	everywhere	inward	normally	secondly	thereby	way
across	between	ex	is	nothing	see	therefore	we
actually	beyond	exactly	it	novel	seeing	therein	we’d
after	both	example	it’d	now	seem	theres	we’ll
afterwards	brief	except	it’ll	nowhere	seemed	thereupon	we’re
again	but	far	it’s	obviously	seeming	these	we’ve
against	by	few	its	of	seems	they	welcome
all	came	fifth	itself	off	seen	they’d	well
allow	can	five	just	often	self	they’ll	went
allows	cannot	followed	know	oh	selves	they’re	were
almost	cant	following	knows	old	sensible	they’ve	what
alone	cause	follows	known	on	sent	think	what’s
along	causes	for	last	once	seven	third	whatever
already	certain	former	lately	one	several	this	when
also	certainly	formerly	later	ones	shall	thorough	whence
although	changes	forth	latter	only	she	thoroughly	whenever
always	clearly	four	latterly	onto	should	those	where
am	co	from	let	or	since	though	where’s
among	com	further	let’s	other	six	three	whereafter
amongst	come	furthermore	like	others	so	through	whereas
an	comes	had	liked	otherwise	some	throughout	whereby
and	concerning	happens	likely	ought	somebody	thru	wherein
another	consequently	hardly	little	our	somehow	thus	whereupon
any	consider	he	look	ours	someone	to	wherever
anybody	considering	hello	looking	ourselves	something	together	whether
anyhow	contain	hence	looks	out	sometime	too	which
anyone	containing	her	ltd	outside	sometimes	took	while
anything	contains	here	mainly	over	somewhat	toward	whither
anyway	corresponding	hereafter	many	overall	somewhere	towards	who
anyways	could	hereby	may	own	soon	tried	who’s
anywhere	course	herein	maybe	particular	sorry	tries	whoever
apart	currently	hereupon	me	particularly	specified	truly	whole
appear	definitely	hers	mean	per	specify	try	whom
appreciate	described	herself	meanwhile	perhaps	specifying	trying	whose
appropriate	despite	hi	merely	placed	still	twice	why
are	did	him	might	please	sub	two	will
around	different	himself	more	plus	such	un	willing
as	do	his	moreover	possible	sup	under	wish
aside	does	hither	most	presumably	sure	unfortunately	with
ask	doing	how	mostly	probably	t	unless	within
asking	done	howbeit	much	provides	take	unlikely	without
associated	down	however	must	que	taken	until	won’t
at	downwards	i	my	quite	tell	unto	wonder
away	each	i’ll	name	rather	th	upon	would
awfully	edu	i’m	namely	rd	than	us	yes
be	eg	i’ve	near	re	thank	use	yet
became	eight	ie	nearly	really	thanks	used	you
because	either	if	necessary	reasonably	thanx	useful	you’d
become	else	ignored	need	regarding	that	uses	you’ll
becomes	elsewhere	immediate	needs	regardless	that’s	using	you’re
becoming	enough	in	neither	regards	the	usually	your
before	especially	inc	nevertheless	respectively	their	value	yours
beforehand	et	indeed	new	right	theirs	various	yourself
behind	etc	indicate	next	said	them	very	yourselves
being	even	indicated	nine	same	themselves	via	zero

Appendix B

Sample data provided by Twitter ($MSFT)

Date	Hour	User Name	Nickname	Bio	Tweet content	RTs	Latitude	Longitude	Country	Followers	Following	Listed	Tweet language (ISO 639-1)	Tweet Url	Is a RT	Original Tweet User Name	User Mentions	Hashtags	Symbols
15.06.2016	12:32	Mike Abramovich	mabramovichx	Quant Trading Strategist & Fixed income Portfolio Manager, C/C++, Java, Mathematica & ALGO TRADER. founder of institute for applied Nonsense	$MSFTthis$26Billion better work otherwise #SatyaNadella will be fired so hard he will regret ever taking this job		51.509	-0.126	GB	518	3678	5	en	http://www.twitter.com/mabramovichx/status/743058484950929409	FALSE			#SatyaNadella	$MSFT
15.06.2016	12:30	Tech Stocks	SAlphaTech	Breaking news and analysis about tech stocks from Seeking Alpha. See all SA Twitter options at https://t.co/nzLmsIKexy	Microsoft Vastly Overpays For LinkedIn https://t.co/aGQ0DQbaUO $LNKD $MSFT					3698	30	211	en	http://www.twitter.com/SAlphaTech/status/743057995412578304	FALSE				$LNKD,$MSFT
15.06.2016	12:29	Dave Nash	djnash33	I've worn many hats at a #SaaS #FinTech firm, but what I enjoy most is building connections, helping others, and improving business. Hope to hear from you!	How much is a user worth? $MSFT paid $60 per $LNKD user. $FB dwarfs them at 198 via @pitchbook https://t.co/rVZOK8bFbb #pe #vc #socialmedia	0	53.958	-1.083	GB	626	844	203	en	http://www.twitter.com/djnash33/status/743057856237326336	FALSE		@PitchBook	#pe,#vc,#socialmedia	$MSFT,$LNKD,$FB
15.06.2016	12:27	Bailey Elizabeth	BaileyEBee	Writer for GabelliConnect and Seeking Alpha. Fordham University senior.	Smart move on $MSFTparttoacquire$LNKD. Great company with a sustainable business model. Good that they aren't making structural changes	0	53.958	-1.083	GB	66	220	1	en	http://www.twitter.com/BaileyEBee/status/743057428934098945	FALSE				$MSFT, $LNKD
15.06.2016	12:25	Ryan Hanson	ryahanson	Investment Professional — UND Alumni — Hockey and sports fanatic — Lover of tennis, golf, guitar, and fitness — News tweets are not necessarily an endorsement.	RT @valuewalk: Microsoft Vastly Overpays For LinkedIn https://t.co/x67tJMXB8H $MSFT $LNKD https://t.co/f45ycw2423	6	44.88969	-93.34995000000001	US	308	1278	29	en	http://www.twitter.com/ryahanson/status/743056947994316800	TRUE	@valuewalk	@valuewalk		$MSFT,$LNKD
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮