Mendel
University
in Brno

MENDELU Working Papers
in Business and Economics

67/2016

Analyzing the correlation between online texts and stock price movements at micro-level using machine learning

František Dařena, Jonáš Petrovský, Jan Žižka, Jan Přichystal

MENDELU Working Papers in Business and Economics
Research Centre
Faculty of Business and Economics
Mendel University in Brno
Zemědělská 1, 613 00 Brno
Czech Republic
http://vyzc.pef.mendelu.cz/en
+420 545 132 605

Citation

Dařena, F., Petrovský, J., Žižka, J., Přichystal, J. (2016). Analyzing the correlation between online texts and stock price movements at micro-level using machine learning. *MENDELU Working Papers in Business and Economics* 67/2016. Mendel University in Brno. Cited from: http://ideas.repec.org/s/men/wpaper.html

Abstract

František Dařena, Jonáš Petrovský, Jan Žižka, Jan Přichystal: **Analyzing the correlation between online texts and stock price movements at micro-level using machine learning**

The paper presents the result of experiments that were designed with the goal of revealing the correlation between texts published in online environments (Yahoo! Finance, Facebook, and Twitter) and changes in stock prices of the corresponding companies at a micro level. The association between lexicon detected sentiment and stock prices movements was not confirmed. It was, however, possible to reveal a dependence with application of machine learning based classification. From the experiments it was obvious that the data preparation procedure had a substantial impact on the results. Thus, different stock prices smoothing, lags between documents' release and related stock price changes, five levels of a minimal stock price change, three different weighting schemes for structured document representation, and six classifiers were studied. It has been shown that at least a part of stock price movements is associated to the texts with a proper combination of these parameters.

Key words

Stock price movements, machine learning, classification, textual documents, sentiment

JEL: G17, C38

Contacts

František Dařena, Department of Informatics, Faculty of Business and Economics, Mendel university in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: frantisek.darena@mendelu.cz.

Acknowledgements

We would like to acknowledge support the Czech Science Foundation, grant No. 16-26353S "Sentiment and its Impact on Stock Markets".

Introduction

A lot of research focuses on incorporating vast amount of human opinions into models of various social and economic phenomena. Data provided by digital mediais very popular also in the field of capital markets where it can help in explainig less rational factors like investors' sentiment or public mood as influential for asset pricing and capital market volatility (Bukovina, 2016).

Most of the past research utilized structured data, which is often objective, to analyze impact of volatile data to business (Groth and Muntermann, 20110. Including other information sources and types, like textual sentiment which represents some kind of subjective information, into, e.g., asset pricing models, provides another perspective and potentially complementary information to quantitative information. This type of information may contain additional hard-to-quantify knowledge (Kearney and Liu, 2014). It can be assumed that the sentiment and mood of the public can influence financial decisions in a similar extent as news as (Bollen, Mao and Zheng, 2011) found that collective mood in Twitter messages correlates to the value of the Dow Jones Industrial Average. Textual data can often not only predict changes but explain the reasons of them, too.

Both types of information, objective and subjective, can be expressed in a textual form in various sources. Objective facts are mostly typical for newspaper articles, scientific papers, annual reports, or other professional texts. On the other hand, texts written by regular people in an informal way, without time and spatial limits, shared with their friends or interest groups often contain certain portion of subjective information. From the texts, useful knowledge might be extracted in a process known as text mining (Feldman and Sanger, 2007). Text mining is a branch of computer science that uses techniques from data mining, information retrieval, machine learning, statistics, natural language processing, and knowledge management (Berry and Kogan, 2010). The greatest potential of text mining applications is in the areas where large quantities of textual data are generated and collected. Text mining involves general tasks such as text categorization, term extraction, single- or multi-document document summarization, clustering, association rules mining, or sentiment analysis (Feldman and Sanger, 2007). At the end of the last century, machine learning gained on its popularity and became a dominant approach to text mining (Sebastiani, 2002).

Advantages of using online resources for decision making support include the timeliness of the information, which is particularly important for investment decisions. On the other hand, the quality of the messages posted in online environments (such as microblogs or discussions in social networks) is generally low. Sentiment contained in internet messages is potentially noisier, less accurate and reliable then sentiment in corporate reports or media articles, because it contains more views from

individual traders. That's why internet postings have been the least frequently studied source of textual sentiment (Kearney and Liu, 2014). Despite all difficulties, content generated by web users has become a widely accepted resource for mining sentiment or opinions regarding different aspects of the public mood (Tumasjan et al., 2011). It has been shown, that a large number of people participating in a content generation process enables creating artifacts that are of equal or superior quality than those made by experts in the respective field (Gottschlich and Hinz, 2014). Messages from millions of people are also unlikely to be biased (Mostafa,2013).

(Wuthrich et al., 1998) investigated whether the content of newspaper articles can predict changes in selected composite indices. Their approach is based on training data from 100 days and a set of more than four hundred phrases provided by a human expert. They achieved the prediction accuracy between 40 and 47% with a great portion of additional outcomes that were only slightly wrong and were able to achieve a trading strategy comparable to or better than human managers.

Schumaker and Chen (2009) studied 484 companies from S&P 500 for one month in 2005. They analyzed the impact of news releases to stock price movements. In their experiments using a Support Vector Machine derivative they achieved 56 to 58 of directional accuracy.

Rao and Srivastava (2012) studied several characteristics of Twitter messages and their relation to stock price movements for 13 stock market indices. They found a strong correlation up to 0.88.

Most of the studies analyzed sentiment primarily at aggregate level, based on the existence of specific words or expressions identified by rules or lexicons. Despite numerous attempts and application areas summarized by Hagenau, Liebman and Neumann (2013), prediction accuracies for the direction of stock prices following the release of corporate financial news rarely exceeded 58%. The same authors achieved accuracy about 76% for one data set by employing a particular combination of advanced feature generation and selection methods together with exogenous market feedback.

In our work we focus on the analysis of information content at the micro level, namely at the level of individual companies. The goal is to determine whether the content of online texts related to a company correlates with movements of prices of stocks related to that company. In this research we combine documents from three different sources, Yahoo! Finance, Facebook, and Twitter collected in a period of about 8 months. From the text mining perspective, these resources are of different quality and their processing is a challenging task. On the other hand, the combination of these resources can mitigate the bias and subjectivity that we would be exposed to by using only one resource. A sentiment lexicon based and machine learning approaches are tested in order to find out whether subjective content or all content play an important role in revealing the document-stock price association.

1 Data used in the experiments

In the experiments, data related to so-called blue chip (large and famous) companies was used. The reason for this choice was a higher probability of availability of sufficient amount of related texts. The analyzed companies were selected from Standard Poor's 500 and FTSEurofirst 300 as they contain a sufficient number of listed companies both US based and European.

In order to analyze the relationship between stock price movements and facts and opinions expressed by internet users, two types of data were needed – stock prices at desired moments in time, and texts containing information related to the selected companies.

The information about stock prices might be obtained at stock exchanges or in specialized internet data sources. For our purpose, Yahoo! Finance was selected as a suitable one as it contains daily data for many stock exchanges around the whole world, with long history, and is available free of charge. For every working day and company, opening, highest, lowest, closing, and adjusted closing stock prices are available together with traded volumes.

Texts related to the investigated companies might be found in many different sources. Usually, the objective ones are typically found on news servers. Texts containing also subjective opinions are usually contained on places where the content is created by individuals without many constraints imposed on the content. These places include social networks, microblogging sites, instant messaging platforms, sites for multimedia sharing, or discussion forums.

From the available financial news servers Yahoo! Finance was selected. It contains news aggregated from several sources (unlike, e.g., Reuters.com), is one of the most visited servers (measured by the Alexa rank), contains also recommendations of financial analysts, and is accessible free of charge.

Other sources of data include social networks and microblogging sites Facebook and Twitter. They belong to the biggest sites on the web, are used in the entire world (are not limited, e.g., to China), provide free public access through their APIs, contain a lot of text data; Twitter also enables searching for a specific content.

On Facebook, companies have their profile pages (fan pages). From the investigated companies, only 55% had such a page. There is a sequence of documents, called posts, arranged according to the time of their publishing in a timeline. These short postings are created by the company representatives. The posts might be commented by other Facebook users at any moment. The comments, however, don't have to be necessarily related to a particular post (e.g., users are just complaining about company products/services).

Twitter is a microblogging site enabling the users to publish short (up to 140 characters) messages, called tweets. Other users might follow their favorite users (i.e., receive their tweets), answer them, or send them new messages. Twitter provides a searching capability with quite a lot of possibilities. In this work, tweets containing the user name of a company (a query contains, e.g., "@google"), mentioning a company (e.g., "Google"), replies to the tweets of a company (e.g., "to: google"), and tweets from the company timeline were used. Because the amount of data on Twitter is very massive, only 21 representative companies from different industries were investigated.

The mentioned data was downloaded according to a predefined schedule. Information about stock prices was downloaded once in every day. Yahoo! Finance articles were downloaded every day. Every day new posts on Facebook profiles were downloaded. Together with them, 100 most liked comments were downloaded, too. Every day, new comments to existing posts and numbers of likings by users were detected. Twitter data was retrieved every six hours because of larger volumes and inability to retrieve more than 100 tweets at a moment. Tab. 1 contains the total and average numbers of data items analyzed in the experiments.

Table 1: Amounts of data from different sources (from 2015-08-01 to 2016-04-04)

Document type	Total number	Daily average/company	Monthly average/company
Yahoo! Finance article	81,519	0.40	12.55
Facebook post	134,941	0.71	21.97
Facebook comment	2,222,362	17.59	545.43
Twitter status	3,887,527	774.29	24,003.00

2 Analyzing the relationship between texts and stock prices

To analyze the relation between stock prices and facts and opinions expressed in text documents, both types of data need to be represented in some quantifiable way, i.e., using variables of a suitable type describing the phenomena. These variables need to express the observed reality in an appropriate way, using values that are well chosen towards a given goal. Very often (especially in the text mining domain) such values are derived from the investigated objects. It means that the original values, in our case stock prices and the content of text documents, need to be transformed to satisfy the requirements of the research and procedures used in it. Ideally, each phenomenon should be represented by a simple variable so the correlation between them can be easily calculated.

Stock prices are represented by single numeric values. The absolute values are, however, not too important for our task as the changes between certain moments in time are. Thus, a transformation based on stock price changes needs to be defined.

Text documents consist of unstructured information in a form of text written in a natural language. The properties of the documents are given by the semantic content defined by the words and their combinations contained in them. The popular bag-of-words approach is based on the principle that every word in the document represents a single variable that together characterize the document. Every document is then descried by a vector of variables where its component values are calculated based on the word frequencies in the document and the entire collection.

In the field of capital markets, behavioral finance considers factors like investors' sentiment or public mood as influential for asset pricing and capital market volatility. Thus, sentiment analysis is one of the important research approaches used in this area in the last few years (Bukovina, 2016). Sentiment analysis mainly studies opinions that express positive or negative sentiments. The most important indicators of sentiment are so called sentiment words or expressions (Liu, 2012) and a comprehensive, high quality lexicon is often essential for fast and accurate sentiment analysis on a large scale (Hutto and Gilbert, 2014}. By application of such a lexicon to a document a single number (e.g., on a scale <-1;+1>) or a nominal value (e.g., negative, neutral, positive) representing the overall sentiment (that represents the document properties) can be determined.

The values representing stock price movements and properties of the related textual documents can be considered time series because every value is associated with a certain date. However, it is not clear when the values of one series react to the values of the other. It can be assumed that the time series are shifted in time relatively to each other which is known as lagged relationship. In this paper we study how financial markets react to news, which is a long lasting question in finance (Wong, Liu and Chiang, 2014). We consider one, two, and three days lags between the documents and stock price movements.

2.1 Handling stock prices

A stock price is represented by a number expressing the price (in, e.g., US dollars) at which the stocks are sold and purchased at a certain moment in time. Because the price is usually volatile (is changing very quickly) during trading periods (in opening hours of a stock exchange), only some of the values are important, especially for historical data. Typically, opening (at the beginning), closing (at the end), low (minimal), and high (maximal) prices in a day are considered (Ang, 2015).

In an investigated period, the stock prices can remain on the same level, which is very rare, or increase or decrease in different rates (slow or rapid). Naturally, the prices change very quickly, although sometimes in small rates, reflecting many different events, habits, or sentiment (Blau and Griffith, 2016). Not all changes are, however, important – after a small drop the price might return to its original

(or higher) level the other day and vice versa, repeating the same movements for a few days or weeks. The price at the end of a week might be thus almost the same as at the beginning while many small movements have been undergone. These movements might have a reason but there is also evidence that price movements might be completely random (Borch, 1963) and it is not necessary to include them in reasoning about the data.

Stock prices can be considered non-stationary time series data and rather trends, cycles, or their combinations are more important (Patel at al., 2014). These movements can be revealed by replacing the original values by some values not showing that high volatility (this process is known as smoothing). The "noise" is replaced which then better represents real and significant price changes. Good candidates are moving averages that substitute the original data by sequences of averages calculated from subsets of the data sets. Changes in these average values are then better indicators of important changes in prices, see Fig. 1.

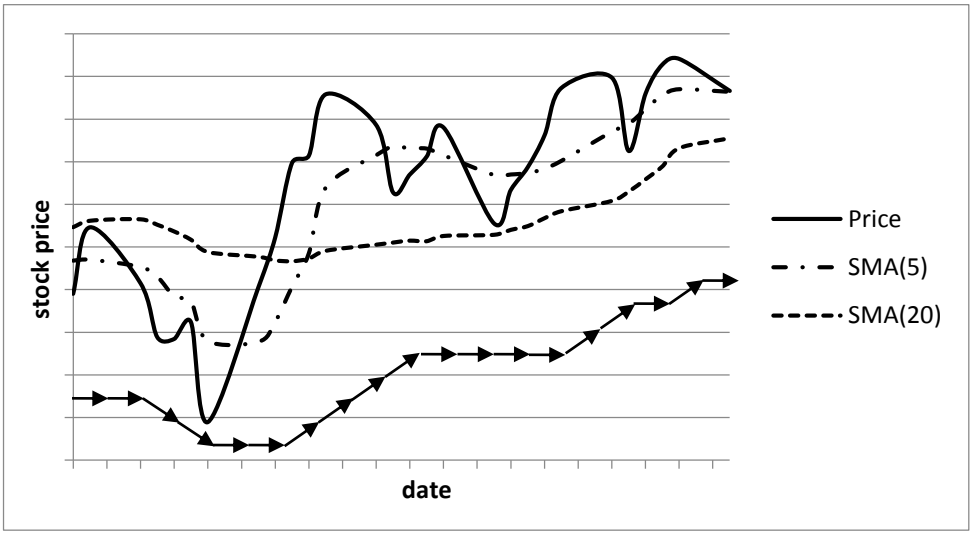

Figure 1: A graph showing stock price development and its smoothing (using Simple Moving Average, SMA, working with 5 and 20 days). The smoothing can better reveal trends in the data as expressed by the arrows. Here, three trend types (increase, stagnation, and decrease) based on a minimal price change are shown for the values smoothed using SMA(5).

Moving averages of different types have been widely used in technical analyses studying stocks markets. Generally, a moving average calculation can work with sequences of subsequent values of different lengths. Short moving averages are more sensitive to changes than long ones (Wang et al., 2014).

Generally, there are two distinct groups of smoothing methods – averaging methods, and exponential smoothing methods, both calculating a new value based on n (here, a number of days) last original

values. The former (Simple Moving Average – SMA) relies on calculating the mean of successive smaller sets of numbers of past data:

$$SMA_t = \frac{price_t + price_{t-1} + \cdots + price_{t-n+1}}{n}$$

The latter (Exponentially Weighted Moving Average – EWMA) assigns exponentially decreasing weights as the observations get older (NIST, 2016):

$$EWMA_t = \lambda \cdot price_t + (1 - \lambda) \cdot EWMA_{t-1}, \lambda = \frac{2}{n+1}$$

In our experiments, besides working with original stock prices, both types of the averages based on two different periods, 5 and 20 days, were considered for calculations in order to include averages with different sensitivities.

At any time, a change that has occurred since the previous moment can be detected. Obviously, very small changes, e.g., in the order of tenths or hundredths of a percent, are usually not important. The question is how big a change needs to be to be considered significant. Wuthrich et al. (1998) found that appreciation and depreciation takes place when the market moves up or down by at least 0.5%. However, the same authors observed that the average change in market indices is often much more, about 1.5%. In our work, the price movements were considered significant if the price changed by 1, 2, 3, 4, or 5 percent. Positive and negative changes above this threshold are then considered price increases and price drops (decreases), respectively. They then represent the classes (categories) for the stock prices data set.

2.2 Handling text data

Text documents generally contain information that has some relationship to reality (the reality is described, evaluated, judged, compared). Understanding the messages might then help with interpreting or predicting events in reality without explicitly observing and studying it. For example, after looking at customer reviews of hotel accommodation at a travellers' website business performance of a hotel might be predicted (Ye, Law and Gu, 2009).

This information consisting of objective facts, personal attitudes, feelings, assumptions, current mood, etc. is expressed by the words and their combinations contained in the text. A perfect understanding of the meaning of a text and its relation to reality is, however, a complicated task often not faultlessly accomplished even by human experts. Nevertheless, for many tasks perfect and complete comprehension of the text is not needed. It is, for example, possible to determine the main topic of a newspaper article on the basis of presence of some keywords in the text. Similarly, according to its

properties (contained words, number of words, text visibility, presence of hyperlinks, sender's address, etc.) an e-mail can be classified as spam or non-spam.

In the last years a lot of research has been devoted to extracting useful knowledge (e.g., sentiment or included topics) from texts written in natural languages. Some of the approaches are based on lexicons and sets of additional rules. The extracted semantic content then depends on the presence of some of the predefined words or expressions from a lexicon, possibly considering more complex issues, like negation, intensification, irrealis blocking, or intra-sentence and inter-sentence conjunctions (Taboada et al., 2011; Cho et al., 2014).

Other approaches rather rely on availability of a sufficient amount of suitable data from which a model can be learned. These data-driven methods use existing data models for which their parameters need to be estimated or an algorithmic approach that tries to find a new function that models the data. The latter approach, often called machine learning, can be successfully used on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets (Breiman, 2001). For many natural language processing tasks, a machine learning approach performs better than a dictionary based approach (Li, 2010).

2.2.1 Using lexicons to derive properties of text documents

The principle of sentiment extraction based on sentiment lexicons is looking for sentimental words or expressions in texts and taking their sentiment categories or orientation into consideration. The sentiment might be expressed on a three-level scale (typically negative, neutral, and positive, or -1, 0, and 1) or on a finer grained scale (e.g., in the range -5 to +5). All occurrences of significant words or expressions and their sentiment values are then averaged, counted, or aggregated in another way. The final decision on the document/sentence/expression sentiment depends on the used scale and on the type of information needed. The decision results might be, for example, that a document is positive on aggregate, or that it contains both positive and negative parts, or that the sum of weights of all positive expressions is x while the sum of weights of all positive expressions is y (Thellwal et al., 2010).

In order to achieve satisfactory results a sufficiently big and high-quality lexicon must be available. The problem is that a word or expression might have different sentiment polarity in different domains. Thus, using a sentiment lexicon, manually or automatically created for one domain does not have to work well in a different domain. There also exist several general (domain independent) sentiment lexicons, see, e.g., (Baccianella, Esuli and Sebastiani, 2010), but they are often ineffective for stock market contents, too (Oliveira, Cortez and Areal, 2016).

There exist many available sentiment lexicons (Petrovský, 2016). It can be noticed that they significantly differ in the number of words or expressions they contain (from a few hundreds to about 150,000). They are also tailored to different domains or are domain independent. Determining what a correct lexicon is, however, depends on a particular task and source of the data used in the research. For analysing texts from microblogging sites a lexicon might be, for example, enriched by including a list of emoticons to increase accuracy of sentiment detection (Arias, Arratia and Xuriguera, 2013).

Using lexicons for sentiment determination is connected to several difficulties negatively affecting the results. Besides domain specificity, they include word sense disambiguation when looking at a particular word in a lexicon (Hung and Chen, 2016), distinguishing between parts of speech when finding sentimental words (Maks and Vossen, 2012), or inability to handle informal expressions that are typical, e.g., for Twitter messages (Bravo-Marquez, Frank and Pfahringer, 2016).

2.2.2 Using machine learning to derive properties of text documents

Textual documents contain mostly unstructured information which is not suitable in terms of effectivity and efficiency for most of knowledge discovery procedures. Texts are therefore usually converted to a representation, typically a structured one, which is more suitable for a particular task. A widely used structured format is the vector space model proposed by Salton and McGill (1983). Every document is represented by a vector where individual dimensions correspond to the features (terms) and the values are the weights (importance) of the features. The weight w_{ij} of every term i in document j is given by three components – a local weight lw_{ij} representing the frequency in every single document, a global weight gw_i reflecting the discriminative ability of the term, based on the distribution of the term in the entire document collection, and a normalization factor n_j correcting the impact of different document lengths (Salton and McGill, 1983). Popular weighing measures include term frequency and term presence for the local weight (Singhal, 1997), inverse document frequency for the global weight (Robertson, 2004), and the cosine normalization (Chisholm and Kolda, 1999) as the normalization factor. All vectors then form so called document-term matrix where the rows represent the documents and the columns correspond to the terms in the documents.

Very often, the features correspond to the words contained in the documents. Such a simple approach, known as the bag-of-words approach, is popular because of its simplicity and straightforward process of creation while providing satisfactory results (Joachims, 2002). The terms might be, however, also multi-word expressions or can be derived from the original document content. Text mining heavily relies on application of various preprocessing techniques including, e.g., text cleaning, white space removal, case folding, spelling errors corrections, abbreviations expanding, stemming, stop words

removal, negation handling, and finally feature selection (Carvalho, de Matos and Rocio, 2007; Clark and araki, 2011; Haddi, Liu and Shi, 2013). These techniques influence what will be the features characterizing the documents.

In order to quantify the relationship between stock prices and related texts a classifier that assigns a label to a text might be trained. The label assignment is abased on the values of attributes derived from the texts. This label should be then correlated to the classes (movement trends) derived from the stock price changes of the corresponding time series.

A classifier implements a function that assigns labels to objects provided on the input. This function h, called the hypothesis, can be induced from existing examples of input-output pairs, known as training examples. The outputs were generated by an unknown function y. The goal of training (it is a supervised learning problem) is to find the hypothesis that well approximates y. This function can be subsequently used for assigning labels to new, unseen instances. In order to test the accuracy of a hypothesis, a testing set of samples, distinct from the training ones, is provided. After comparing the values of y and h for the testing samples, we can find out how well the hypothesis generalizes. When the values of y are discrete, the process is known as classification (Russel and Norwig, 2010).

For the training phase, a sufficient amount of training instances needed to be prepared and appropriately labeled. For every particular text, the date of its publication and a related company was known. It was then possible to take the stock price movement trend for that company for a corresponding date (considering also a lag) and use it as a label for the document. The induced classifier than tried to map the document features to the labels derived from stock price movements.

To measure the quality of the trained classifiers, i.e., their ability to be used acceptably in the future, they are examined on test samples. The values representing correctly and incorrectly classified examples are used to compute measures of classifier effectiveness. In the two class classification, the classes might be labeled as positive and negative. The positive and negative examples that are classified correctly are referred as true positive (TP) and true negative (TN), respectively. False positive (FP) and false negative (FN) represent misclassified positive and negative examples. Commonly accepted classifier performance evaluation measures include accuracy, precision, recall, and F-measure combining the values of TP, TN, FP, and FN into a single measure (Sokolova, Japkowicz and Szpakowicz, 2006).

The strength of the relationship between the input (the content of documents) and output (the label) might be then expressed by standard classification performance measures, like the accuracy of F-measure since they contain information how well is a classifier able to assign a correct label to a

document based on the values of its attributes. High values of these measures say that there exist attributes or their combinations that are accurately able to distinguish between instances of different classes.

3 Experiments

Four different data sources (newspaper articles, Facebook posts and comments, and tweets) were investigated separately. An amount of documents manageable from the computational complexity perspective needed to be selected. Thus, only 200 most retweeted tweets and 40 most liked comments for every company in every day were processed. The amount of the two remaining data sets, i.e., Facebook posts and Yahoo! Finance articles, were not that high so no preselection needed to be performed.

For both the lexicon based and machine learning based approaches the stock price time series needed to be transformed according to the principle described above. For the machine learning based procedure, a suitable class label for training a classifier in order to determine the correlation with stock price movements needed to be assigned to every text. In order to transform the stock price data and to determine the class label of a document D_i related to company C_i, released at time T_r, representing a change in stock price of company C_i at time T_c the following aspects and parameters need to be determined:

- Concrete values of stock prices to be considered – here, adjusted closing values, simple moving average and exponential moving average, both based on 5 and 20 days were analyzed; for days when no value was available (weekends, holidays), the price was calculated as the arithmetic average of the last closing value and the first following opening value.
- The lag between publication of texts at date T_r and a stock price movement at T_c – lags of 1, 2, and 3 days were investigated.
- The minimal relative difference in stock prices at T_c and T_{c-1} to be considered significant – changes 1, 2, 3, 4, and 5 percent were investigated. If a price change is within the percentage limits it is considered constant and all documents related to the specific date are labeled by the “constant” class label. If the price change is above the limit in the positive direction, i.e., increased more than, e.g., 3%, documents are labeled as “increase”. In the remaining case, the price decreased significantly and the corresponding documents are labeled by the “decrease” label.

As the data was massively unbalanced (a large majority of documents belonged to days when no significant change in stock prices occurred and were thus labeled as “constant”) biased or useless results in terms of accuracy would be achieved without further data set adjustment. Because significant increases or decreases of prices are more interesting than remaining approximately on the same level, documents labeled as “constant” were excluded from further processing and the correlation between texts and stock price movements was analyzed only in periods with significant price changes.

With the prepared data two different types of experiments were carried out. They are described in the following sections.

3.1 Using lexicons to estimate stock price movements

As one can expect, documents containing positive sentiment about a company should be connected to stock price increase. On the contrary, stock price drop should accompany negative sentiment. For this kind of analysis, we now have two variables – sentiment contained in text documents revealed using a sentiment lexicon and categories derived from stock prices changes. To make the quantification of the correlation between them comparable to the other experiments (machine learning based procedure) the same set of metrics was used. In fact, sentiment in a document (or a document collection) can be considered a factor assigning a direction (class) to a stock price movement (positive sentiment = increase, negative sentiment = decrease, and neutral sentiment = constant). The actual movement should be, in an ideal case, the same as the predicted movement which can be measured using standard classification performance measures, like accuracy, precision, recall, or F-measure.

To determine the sentiment contained in the investigated texts the VADER algorithm was used. The algorithm enables determining the compound sentiment of a given piece of text (typically a sentence, but might by applied to whole documents) with using a manually created sentiment lexicon with five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. The model is especially attuned to microblog-like contexts and demonstrates great correlation with judgements of humans (Hutto and Gilbert, 2014).

The output of the VADER algorithm is a number from [-1; 1] scale representing sentiment polarity. To determine a particular sentiment class, e.g., negative, neutral, and positive, some thresholds for the sentiment value need to be specified. Similarly to Hutto and Gilbert (2014), these thresholds were set to the values -0.05 and +0.05.

Considering all possible parameters of this procedure, i.e., five options for stock price values transformation (adjusted close, simple and exponential moving averages working with 5 and 20 days),

three options for the lag (1, 2, or 3 days), and five options for class determination (change 1-5 percent), 75 data sets where the expected document class was determined differently were prepared. These class labels were then compared to the outputs of VADER and the necessary metrics for measuring the success of the process were calculated. To make the experiments comparable to the machine learning based experiments only positive and negative classes were considered.

3.2 Analyzing the dependence between stock prices and texts using classification

The texts of documents were modified in the way that all HTML tags, @ and # characters (marking user names and hashtags on Twitter) and other non-alphanumeric characters were removed, selected emoticons were replaced by artificial terms representing positive and negative sentiment, all URLs were replaced by a single artificial term, and the text was converted to lower case. The minimal length of processed words was 2, and the minimal document frequency of terms was 10 for Yahoo! Finance articled and 5 for the other collections. The texts were converted to vectors using the bag-of-word approach to become acceptable for machine learning algorithms. As weighting schemes, three possibilities were investigated – simple term presence, term frequency with the inverse document frequency weight (tf-idf), and tf-idf with cosine normalization.

In order to not bias a classifier against one bigger class the numbers of documents from both classes ("increase" and "decrease") were balanced.

From the great amount of existing classifiers, the following ones, available in Python's scikit-learn package (Lorica, 2015) were investigated: Multinomial Naïve Bayes (with α=1, i.e., Laplace smoothing), Bernoulli Naïve Bayes, Logistic regression (Maximum entropy), CART decision tree, Random forest, and Linear SVC (Support vector machine with a linear kernel). These algorithms belong to the ones often used in sentiment analysis and text classification (Pang, Lee and Vaithyanathan, 2002; Žižka and Dařena, 2015a). The data was split into training and testing sets in the proportion 65:35 percent.

To make the experiment's results comparable to the lexicon based approach the same methods for document class determination and stock price series transformation were used. Seventy-five different data sets containing documents labelled differently were then encoded using the three weighing schemes (term presence, tf-idf, and tf-idf with cosine normalization) into three different representations which were later supplied to the chosen sclassifiers.

4 Results and discussion

4.1 Lexicon based analysis

All documents related to particular companies were, based on their content, labeled as positive, neutral, or negative using the sentiment lexicon and algorithm described above. When processing Yahoo! Finance articles, sentiment calculation was based on the aggregation of sentiment at sentence level as the VADER algorithm is tuned to work with sentences. The overall sentiment for a particular company and day was then calculated as the prevailing sentiment for all texts related to the company released in that day.

The number of days with positive aggregate sentiment generally largely exceeded the number of days with negative sentiment, in a ratio of 5:1 to 20:1, depending on the document source. On the contrary, the number of days in positive and negative classes, based on price movements was mostly in a ratio of 1:1 to 1:2 for the settings with a sufficient amount of available data. The results of comparing actual classes (based on stock price movements) with predicted classes (based on sentiment) were thus strongly biased towards the positive class. Accuracy was therefore not an ideal performance measure. For that reason, the presented results contain the values of F-measure, too.

The classes (for each company and day) predicted with sentiment analysis were compared to the classes based on all combinations (75 in total) of stock price change category determination parameters, i.e., combinations of a smoothing method, minimal price change, and lag in days. The correctness of the matches between these two values was aggregated and 75 sets of classification performance measure values for each data source were obtained. These values were then averaged with a simple arithmetic average and a weighted average using the numbers of processed items in the experiments as the weights (the results of experiments with a higher number of items had a higher weight). The aggregated values, from the perspective of the three variable parameters, are presented in Tab. 2. As the differences between the values obtained for each of the four data sources were not significant the results aggregated over all experiments are presented.

The smoothing method and minimal price change influenced the amount of data available for experiments. Higher numbers of days used for smoothing and higher minimal price change decreased the numbers of available items. Generally, when only tens of data items were available the values of accuracy or F-measure quantifying the results were lower than in the case of experiments with thousands or tens of thousands of items.

Table 2: Aggregate values of accuracy and F-measure representing the correlation between stock price movement and sentiment of related documents.

	Accuracy		F-measure	
	Average	Weighted average	Average	Weighted average
Smoothing method				
adjclose	0.4623	0.4920	0.3890	0.4018
sma(5)	0.3671	0.4561	0.3301	0.3900
sma(20)	0.2211	0.3028	0.2083	0.2914
ewma(5)	0.3518	0.4413	0.3205	0.3803
ewma(20)	0.2181	0.2994	0.2127	0.2894
Minimal price change				
1%	0.4241	0.4820	0.3700	0.3982
2%	0.3528	0.4591	0.3174	0.3861
3%	0.3086	0.4334	0.2820	0.3722
4%	0.2717	0.4148	0.2486	0.3617
5%	0.2631	0.3836	0.2425	0.3429
Lag in days				
1	0.3355	0.4710	0.3029	0.3936
2	0.3251	0.4680	0.2930	0.3913
3	0.3117	0.4649	0.2804	0.3873

The correctness of the proposed approach is generally quite low, with accuracy and F-measure values below 0.5, decreasing with the decreasing number of data items available for the experiments. The influence of the smoothing method and minimal price change parameters cannot be thus reliably determined. The only parameter for which comparable data exist was the lag in days. Here, the highest values of performance measures can be identified for the value of 1 day.

4.2 Classification based analysis

The data collections for experimenting were prepared according to the steps described in the previous sections. Subsequently, six different classifiers were trained and tested on each of the data sets represented by three different term weighting schemes. Values of the metrics related to classification correctness were obtained for each experiment. To achieve sufficiently general results, collections with less than 500 documents were excluded from detailed analyses of the experiments with individual data sources.

Selected statistical measures of the most important classification performance metrics and data set properties for all experiments can be found in Tab. 3. The values here are based on experiments using all possible combinations of parameters. Because the collections were almost perfectly balanced in

terms of class distribution in the data sets, the values of accuracy, precision, recall, and F-measure reached almost the same values. Thus, in the following text, only the values of accuracy are presented.

Table 3: Classification performance metrics values and data set characteristics for all experiments with data from all four sources.

	Average accuracy	Min. accuracy	Max. accuracy	Accuracy variance	Average number of documents	Average number of attributes
Yahoo! Finance articles	0,6376	0,5430	0,8142	0,0025	10911	13597
Facebook posts	0,5819	0,5020	0,6940	0,0013	14191	6743
Facebook comments	0,6039	0,5229	0,7861	0,0027	43037	10456
Twitter	0,6658	0,5531	0,8393	0,0022	35768	8459

From the Tab. 2 it is obvious that the accuracy varies quite significantly from its minimal to maximal values which is given by different experimental settings. In practice, the experiments where higher accuracies are achieved are more interesting. Thus, a detailed exploration of the used algorithms and experimental settings was conducted in order to reveal how individual parameters influence the success of the classification process. For every variable parameter (a method of stock price values smoothing, a lag between documents' release and related stock price changes, minimal stock price change, classifier, and weighting scheme) average accuracies for all experiments with the same value of the parameter were calculated in order to reveal whether some parameter values lead to better results. The achieved average accuracies can be found in Tab. 4.

From the Tab. 4 it is obvious that only the smoothing method and used classifier had a significant impact on accuracy values. Higher accuracies were achieved for sma(20) and ewma(20) and for LinearSVC, MaxEnt, and NB-multi classifiers across all data sources (the average accuracies for all combinations containing only these values for respective parameters increased to 0.72 for Yahoo! Finance articles, 0.61 for Facebook posts, 0.67 for Facebook comments, and 0.70 for Twitter). For further analysis, only these parameter values were considered to better evaluate the impact of the remaining parameters.

When bigger minimal stock price changes were considered in the experiments, achieved accuracies had a tendency to be higher. From the used parameters, the minimal stock price difference was the parameter that influenced the size of data set the most. The higher was the minimal change to be considered significant, the smaller number of documents labeled as "increase" or "decrease" were available. The experiments were thus carried out with different numbers of documents based on the

value of the minimal stock price change parameter. In order to take this into consideration when looking at the result of subsequent analyses not only average accuracies, but also average accuracies weighted by the number of documents used in the experiments were calculated. The values of both achieved accuracies are presented in Tab. 5.

Table 4: Average accuracies for individual experiments' parameters.

Lag in days	1	2	3
Yahoo! Finance articles	0.6374	0.6345	0.6405
Facebook posts	0.6006	0.5756	0.5729
Facebook comments	0.6025	0.6090	0.6003
Twitter	0.6439	0.6743	0.6754

Minimal price change	1%	2%	3%	4%	5%
Yahoo! Finance articles	0.6339	0.6436	0.6375	0.6413	0.6311
Facebook posts	0.5759	0.5825	0.5823	0.5809	0.5892
Facebook comments	0.5766	0.6068	0.6080	0.6162	0.6178
Twitter	0.6654	0.6795	0.6785	0.6537	0.6456

Smoothing method	adjclose	sma(5)	ewma(5)	sma(20)	ewma(20)
Yahoo! Finance articles	0.6046	0.6239	0.6162	0.6866	0.6903
Facebook posts	0.5921	0.5531	0.5713	0.5980	0.6018
Facebook comments	0.5533	0.5910	0.5944	0.6531	0.6544
Twitter	0.6313	0.6582	0.6657	0.7009	0.6848

Document representation	tf-idf-cos	tf-idf-no	tp-no-no
Yahoo! Finance articles	0.6338	0.6412	0.6378
Facebook posts	0.5810	0.5824	0.5822
Facebook comments	0.6035	0.6043	0.6040
Twitter	0.6609	0.6681	0.6683

Classifier	CART	LinearSVC	MaxEnt	NB-berno	NB-multi	RandForest
Yahoo! Finance articles	0.6089	0.6625	0.6600	0.6232	0.6507	0.6202
Facebook posts	0.5585	0.5798	0.5868	0.5986	0.5967	0.5708
Facebook comments	0.5839	0.6086	0.6133	0.6083	0.6154	0.5940
Twitter	0.6512	0.6724	0.6683	0.6667	0.6644	0.6717

Because of high volatility of the stock price data smoothing of the time series has proven to be a reasonable step improving the accuracy for most of the data sources significantly. Moving averages based on 20 days had more positive impact than moving averages based on 5 days. The type of moving average (simple or exponential) was not considerably important.

Table 5: Average accuracies (AVG) and weighted average accuracies (WAVG) for individual experiments' parameters. All experiments with classifiers and smoothing methods different than shown were excluded.

Lag in days	1		2		3	
	AVG	WAVG	AVG	WAVG	AVG	WAVG
Yahoo! Finance	0,7414	0,7321	0,7181	0,6980	0,7125	0,6835
Facebook comments	0,6711	0,6765	0,6752	0,6345	0,6552	0,6117
Facebook posts	0,6449	0,6372	0,5980	0,5923	0,5960	0,5811
Twitter	0,6761	0,6854	0,6992	0,7317	0,7045	0,7042

Minimal price change	1		2		3		4		5	
	AVG	WAVG	AVG	WAVG	AVG	WAVG	AVG	WAVG	AVG	WAVG
Yahoo! Finance	0,6958	0,6786	0,7329	0,7232	0,7142	0,7168	0,7283	0,7126	0,7474	0,7483
Facebook comments	0,6291	0,6101	0,6632	0,6565	0,6677	0,6633	0,6858	0,6758	0,7187	0,6930
Facebook posts	0,6007	0,5839	0,6074	0,5973	0,6001	0,5851	0,6028	0,6001	0,6356	0,6441
Twitter	0,7137	0,7104	0,7132	0,7480	0,7044	0,6897	0,6542	0,6468	0,6598	0,6421

Classifier	LinearSVC		MaxEnt		NB-multi	
	AVG	WAVG	AVG	WAVG	AVG	WAVG
Yahoo! Finance	0,7250	0,6986	0,7210	0,6997	0,7104	0,6726
Facebook comments	0,6589	0,6210	0,6673	0,6261	0,6700	0,6244
Facebook posts	0,5993	0,5804	0,6064	0,5855	0,6151	0,5981
Twitter	0,7015	0,7174	0,6954	0,7136	0,6926	0,7047

Document representation	tf-idf-cos		tf-idf-no		tp-no-no	
	AVG	WAVG	AVG	WAVG	AVG	WAVG
Yahoo! Finance	0,7106	0,6884	0,7284	0,6957	0,7173	0,6868
Facebook comments	0,6676	0,6254	0,6648	0,6232	0,6637	0,6230
Facebook posts	0,5817	0,5893	0,5847	0,5887	0,5852	0,5860
Twitter	0,6905	0,7069	0,6997	0,7138	0,6992	0,7151

Smoothing method	sma(20)		ewma(20)	
	AVG	WAVG	AVG	WAVG
Yahoo! Finance	0,7203	0,6972	0,7176	0,6843
Facebook comments	0,6651	0,6244	0,6656	0,6233
Facebook posts	0,6065	0,5912	0,6073	0,5856
Twitter	0,7018	0,7163	0,6906	0,7080

When looking at the time between publication of documents and related stock price changes the strongest correlation was found for shorter time spans for the Yahoo! Finance and Facebook documents (1 day, or 1-2 days, respectively) and longer (2-3 days) for Twitter. It can be thus seen that the content of the documents correlated with stock price movements differently distant from their

publication according to the document source. A possible explanation might be in the nature of the documents. As it takes some time to publish a newspaper article, the time distance between an article and a price movement is rather short. Texts that are published very quickly, like Twitter messages, might anticipate a price movement earlier. Facebook posts that are often prepared by company representatives are usually not published timely so their nature is in this respect more similar to newspaper articles. The comments created by other people are sometimes immediate, sometimes delayed.

For all data sources, except Twitter, higher considered minimal stock price changes lead to better results in terms of classification accuracy. We can assume that these substantial changes were accompanied by an exceptional content of documents that makes them more distinguishable from the documents published in periods with no or small price changes. This parameter, however, influences the size of available data (there are less periods with big changes that periods with small changes) so the possibility of mining useful knowledge from the data might be limited.

The impact of different weighting methods was very low; the average accuracies lie in an interval of about 1%. Thus, the weighting scheme can be considered an unimportant factor of data preprocessing.

5 Conclusion

The paper presents the result of experiments that were designed with the goal of revealing the correlation between texts published in online environments (Yahoo! Finance articles, Facebook posts and comments, and Twitter messages) and changes in stock prices of the corresponding companies at a micro level.

The association between sentiment (detected with application of a sentiment lexicon) contained in the documents and movements of stock prices was not confirmed. The correlation expressed by correctness of matching positive sentiment to stock price increase and negative sentiment to stock price decrease was very low as measured by the accuracy and F-measure.

It was, however, possible to reveal a dependence between texts published in newspapers and on social networks and microblogging sites with application of machine learning based classification. Here, also other that subjective content played a significant role and could be used to distinguish between positive and negative stock price movements.

All used classifiers were able to confirm the correlation between texts and stock price movements with all data sets prepared for the conducted experiments. Some of them, namely Linear SVC, Maximum

Entropy, and multinomial Naïve Bayes classifiers outperformed the others in terms of achieved accuracy (which was, however, not the mail research goal). The difference between the maximal and minimal achieved accuracies for the same data was from about 20 to 30%. It was therefore obvious that the data preparation procedure had a substantial impact on the results.

To make the correlation quantifiable several methods of transformation of the two time-series (texts and stock prices) were carried out. Stock prices were smoothed by four different methods, three different lags between documents' release and related stock price changes were considered, five levels of a minimal stock price change to take the change for significant were used, and three different weighting schemes for structured document representation were examined. From these parameters, the smoothing method played the most important role. It was found that smoothing the stock price data with moving averages based on 20 preceding days led to better results than in case of using only 5 days. Such smoothing removed excessive price oscillations which are quite typical for this type of data and are often random. On the other hand, some of the important changes, especially when followed by another change in the opposite direction might be lost.

There are many aspects that influence stock price movements and that are not always included in the online texts. It is thus clear that the documents' content cannot explain or predict all movements. It has been shown that at least a part of these movements is associated to the texts and can be used as a part of a more complex model of an economic phenomenon.

Future research directions will include a tighter interconnection with the economic aspects of the domain, including, e.g., other external market and economy information and industry specifics. A special attention will be paid to the process of transformation of texts to their structured representation including specific approaches to processing texts from different data sources and their combinations. From the machine learning perspective, processing the data in a stream using, e.g., a moving window approach (Žižka and Dařena, 2015b), processing unbalanced data, or including additional features like the dynamics of Facebook posts and comments likings or Yahoo! Finance articles sharing.

References

ANG, C., 2015: Analyzing Financial Data and Implementing Financial Models Using R. Springer.

ARIAS, M., ARRATIA, A., XURIGUERA, R., 2013: Forecasting with Twitter Data, ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 1, pp. 8:1–8:24.

BACCIANELLA, S., ESULI, A., SEBASTIANI, F., 2010: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation LREC10, European Language Resources Association (ELRA), pp. 2200–2204.

BLAU, B. M., GRIFFITH, T. G., 2016: Price clustering and the stability of stock prices. Journal of Business Research, vol. 69, no. 10, pp. 3933–3942.

BERRY, M. W., KOGAN, J., 2010: Text Mining: Applications and Theory. Chichester: Wiley.

BOLLEN, J., MAO, H., ZENG, X., 2011: Twitter mood predicts the stock market. Journal of Computational Science, vol. 2, no. 1, pp. 1–8.

BORCH, K., 1963: Price movements in the stock market, Econometric research program, research paper no. 7, Princeton University.

BRAVO-MARQUEZ, F., FRANK, E., PFAHRINGER, B., 2016: Building a Twitter opinion lexicon from automatically-annotated tweets. Knowledge-Based Systems, vol. 108, pp. 65–78.

BREIMAN, L., 2001: Statistical modeling: The two cultures (with comments and a rejoinder by the author. Statistical Science, vol. 16, no. 3, pp. 199–231.

BUKOVINA, J., 2016: Social media big data and capital markets—An overview. Journal of Behavioral and Experimental Finance, vol. 11, pp. 18–26.

CARVALHO, G., DE MATOS, D. M., ROCIO, V., 2007: Document retrieval for question answering: a quantitative evaluation of text preprocessing. In Proceedings of the ACM first Ph. D. workshop in CIKM, pp. 125–130.

CHISHOLM, E., KOLDA, T. G., 1999: New term weighting formulas for the vector space method in information retrieval. Computer Science and Mathematics Division, Oak Ridge National Laboratory.

CHO, H., KIM, S., LEE, J., LEE, J. S., 2014: Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. Knowledge-Based Systems, vol. 71, pp. 61–71.

CLARK, E, ARAKI, K., 2011: Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia – Social and Behavioral Sciences, vol. 27, pp. 2–11.

FELDMAN, R., SANGER, J., 2007: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

GOTTSCHLICH, J., HINZ, O., 2014: A decision support system for stock investment recommendations using collective wisdom. Decision Support Systems, vol. 59, pp. 52–62.

GROTH, S. S., MUNTERMANN, J., 2011: An intraday market risk management approach based on textual analysis. Decision Support Systems, vol. 50, no. 4, pp. 680–691.

HADDI, E., LIU, X., SHI, Y., 2013: The role of text pre-processing in sentiment analysis. Procedia Computer Science, vol. 17, pp. 26–32.

HAGENAU, M., LIEBMANN, M., NEUMANN, D., 2013: Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems, vol. 55, no. 3, pp. 685–697.

HUNG, C., CHEN, S. J., 2013: Word sense disambiguation based sentiment lexicons for sentiment classification. Decision Support Systems, vol. 55, no. 3, pp. 685–697.

HUTTO, C. J., GILBERT, E., 2014: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International AAAI Conference on Weblogs and Social Media.

JOACHIMS, T., 2002: Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers.

KEARNEY, C., LIU, S., 2014: Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, vol. 33, pp. 171–185.

LI, F., 2010: The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, vol., 48, no. 5, pp. 1049–1102.

LIU, B., 2012: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167.

LORICA, B., 2015: Six reasons why I recommend scikit-learn, Retrieved November 13, 2016, from https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn.

MAKS, I., VOSSEN, P., 2012: A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems, vol. 53, no. 4, pp. 680–688.

MITCHELL, C., 2016: How to use a moving average to buy stocks — Investopedia. Retrieved May 17, 2016, from http://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp.

MOSTAFA, M.M., 2013), More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications, vol. 40, no. 10, pp. 4241–4251.

NIST/SEMATECH (2016), e-Handbook of Statistical Methods, Retrieved August 11, 2016, from http://www.itl.nist.gov/div898/handbook/.

OLIVEIRA, N., CORTEZ, P., AREAL, N., 2016: Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decision Support Systems, vol. 85, pp. 62–73.

PANG, B., LEE, L., VAITHYANATHAN, S., 2002, July: Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79–86.

PATEL, J., SHAH, S., THAKKAR, P., KOTECHA, K., 2015: Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, vol. 42, no. 1, pp. 259–268.

RAO, T., SRIVASTAVA, S., 2014: Twitter sentiment analysis: How to hedge your bets in the stock markets. In State of the Art Applications of Social Network Analysis, pp. 227–247.

ROBERTSON, S., 2004: Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation, vol. 60, no. 5, pp. 503–520.

RUSSEL, S., NORWIG, P., 2016: Artificial Intelligence: A Modern Approach, Pearson Education, Upper Saddle River.

SALTON, G., MCGILL, M. J., 1983: Introduction to Modern Information Retrieval, McGraw Hill.

SCHUMAKER, R. P., CHEN, H., 2009: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), vol. 27, no. 2.

SINGHAL, A. K., 1997: Term Weighting Revisited, PhD dissertation, Faculty of the Graduate School of Cornell University.

SOKOLOVA, M., JAPKOWICZ, N., SZPAKOWICZ, S., 2006, December: Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian Joint Conference on Artificial Intelligence, pp. 1015–1021.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K., STEDE, M., 2011: Lexicon-based methods for sentiment analysis. Computational linguistics, vol. 37, no. 2, pp 267–307.

THELWALL, M., BUCKLEY, K., PALTOGLOU, G. CAI, D., KAPPAS, A., 2010: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544–2558.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., WELPE, I. M., 2011: Election forecasts with Twitter: How 140 characters reflect the political landscape. Social Science Computer Review, vol. 29, no. 402–418.

WANG, L., AN, H., XIA, X., LIU, X., SUN, X., HUANG, X., 2014: Generating moving average trading rules on the oil futures market with genetic algorithms. Mathematical Problems in Engineering, 2014.

WONG, F. M. F., LIU, Z., CHIANG, M., 2014, December: Stock market prediction from WSJ: text mining via sparse matrix factorization. In 2014 IEEE International Conference on Data Mining, pp. 430–439.

WUTHRICH, B., CHO, V., LEUNG, S., PERMUNETILLEKE, D., SANKARAN, K., ZHANG, J., 1998, October: Daily stock market forecast from textual web data. In Systems, Man, and Cybernetics, 1998, vol. 3, pp. 2720–2725.

YE, Q., LAW, R., GU, B., 2009: The impact of online user reviews on hotel room sales. International Journal of Hospitality Management, vol. 28, no. 1, pp. 180–182.

ŽIŽKA, J., DAŘENA, F., 2015a: Automated Mining of Relevant N-grams in Relation to Predominant Topics of Text Documents. In International Conference on Text, Speech, and Dialogue, pp. 461–469.

ŽIŽKA, J., DAŘENA, F., 2015b: Revealing potential changes of significant terms in streams of textual data written in natural languages using windowing and text mining. In Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 131–138.