ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume 66 142 Number 6, 2018

ANALYSIS OF THE ASSOCIATION BETWEEN TOPICS IN ONLINE DOCUMENTS AND STOCK PRICE MOVEMENTS

František Dařena[1], Jan Přichystal[1]

[1]Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 61300 Brno, Czech Republic

To cite this article: DAŘENA FRANTIŠEK, PŘICHYSTAL JAN. 2018. Analysis of the Association between Topics in Online Documents and Stock Price Movements. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 66(6): 1431–1439.

To link to this article: https://doi.org/10.11118/actaun201866061431

Abstract

This paper aims at discovering the topics hidden in the newspaper articles that have an impact on movements of stock prices of the corresponding companies. Document topics are characterized by combinations of specific words in documents and are shared across a document collection. We describe the process of discovering the topics, the creation of a mapping of the topics to stock price movements, and quantifying and evaluating the results. As the method for finding and quantifying the association, we use machine learning-based classification. We achieved an accuracy of stock price movement predictions higher than 70%. A feature selection procedure was applied to the features characterizing the topics in order to facilitate the process of assigning a label to the topic by a human expert.

Keywords: stock prices, topics in document collections, machine learning, classification, feature selection

INTRODUCTION

A lot of research has been focusing on incorporating data available online into models of various social and economic phenomena. One such domain is the field of capital markets where the data provided by digital media can help in explaining less rational investors' decisions (Bukovina, 2016).

There exist several commercial financial expert systems that can be successfully used for trading on a stock exchange. The ones that rely primarily on time-series analysis of the market have limited capabilities (Weng, Ahmed and Megahed, 2017). It is therefore desirable to include also other information to an investment decision making process, like unstructured texts, published by different types of subjects, containing additional knowledge (Kearney and Liu, 2014). This is demonstrated by Lee *et al.* (2014) that developed a stock price forecasting system combining financial and textual information. In the financial forecasting domain, data mining, text mining, and natural language processing are commonly used disciplines (Kumar and Ravi, 2016).

Some authors, like Schumaker and Chen (2009) examined 484 corporations from the S and P 500 for one month in 2005. They investigated the effect of news on stock price movements. In their experiments they used a Support Vector Machine derivative and bag of words, noun phrases, and named entities representations of texts. Wuthrich *et al.* (1998) studied whether the content of newspaper articles can foresee changes in selected composite indices. Their used training data from 100 days and a set of more than four hundred phrases provided by a human expert.

Some works use sentiment as an indicator of stock price movements. However, focusing simply on the sentiment (positive and negative) dimensions does not always bring useful predictions (Li *et al.*, 2014). In our previous research (Dařena *et al.*, 2018), we revealed that it was possible to uncover and quantify a relationship between unstructured texts and stock price movements at a micro level with the application of machine learning. The data to be classified included the texts converted to their structured representation (bag-of-words). The classes to which the texts were assigned were derived from relative changes in stock prices.

In this paper, we analyze the data on a micro level (the level of individual companies) and do not use any external knowledge (like dictionaries prepared by human experts). We extend our previous research and instead of focusing on the connection between specific words included in online texts and stock price movements we aim at discovering the topics that have an impact on stock price movements. We hypothesize that a topic, characterized by a combination of specific words, can be a good indicator of a stock price movement and can explain specific stock price movements. In current research, different sources of text data, like newspapers, Twitter, Facebook, 8-K forms, or 10-K forms are investigated (Lee *et al.*, 2014; Loughran and McDonald, 2011; Ranco *et al.*, 2015; Siganos, Vagenas-Nanos and Verwijmeren, 2017). In our work, we initially focus on newspaper articles because topic discovery in collections of short informal texts is still an unsolved problem (Zuo *et al.*, 2016).

The paper describes the process of discovering topics in a newspaper articles collection, creating a mapping of the topics to stock price movements, and quantifying and evaluating the results.

MATERIALS AND METHODS

There are two variables between which a relationship should be discovered and quantified – stock prices and texts. We analyzed the companies from Standard and Poor's 500 and FTSEurofirst 300 indices. The information about stock prices was obtained from Yahoo! Finance. The source provides daily data from many stock exchanges in the world free of charge. Textual documents associated with the studied companies from the period from 2014-02-20 to 2016-10-12 were downloaded from Yahoo! Finance. Here, news aggregated from several sources and associated to every company could be found.

A price is a target (dependent) variable with continuous values. The texts represent independent variables having an unstructured form. In many situations, an association between two variables and its strength can be expressed using the Pearson's correlation coefficient (Benesty *et al.*, 2009). When the value of the coefficient equals to 1 the values of both variables move in the same direction (i.e., they both increase or decrease). In order to calculate the value of the coefficient, both values must be scalar. Because texts are, in their original form, not represented by scalar values, such a calculation is not feasible.

Another possibility to describe an association between two variables is a function mapping the values of one variable to another. Thus, the texts need to be transformed to a different form where such a mapping would be possible. A common format used in the text mining domain is the vector space model proposed by Salton and McGill (1983). A vector where individual dimensions correspond to the document features and the values are the importance of the features represents every document. Very often, the features correspond to the words contained in the documents. Such a simple approach is then known as the bag-of-words approach (Joachims, 2002).

Mapping the document vectors to the values of stock prices might be related to many problems. The number of attributes describing texts is generally very high (can be in the order of tens of thousands) and the attribute vectors are sparse (Žižka and Dařena, 2013). A certain type of event can be also characterized by many different words. Their combinations, however, might be important too. Thus, instead of working with isolated words from the documents, topics prevailing in each document might be indicators of the subsequent stock price movements. It is thus necessary to derive a topic (which is not known in advance) from each document.

Processing documents

Clustering, which is the most prevalent task belonging to the family of unsupervised learning methods, enables automatic organization of unlabeled documents into groups called clusters. The clusters are expected to be homogeneous inside and must be clearly distinguishable from each other to express their own distinct information. Clustering has been successfully applied for organizing and searching large text collections (Bsoul *et al.*, 2013; Dhillon and Modha, 1999; Guo and Zhang, 2009; Tseng *et al.*, 2007) and can be used to discover main topics hidden in collections of texts (Žižka and Dařena, 2013; Barák *et al.*, 2015).

In the preprocessing phase, the documents were converted to their vector representations. Rare words, in this case words that appeared five times and less in the collection, were removed. Such rare words (very often just errors) represent the noise in the data negatively influencing the quality of results. Such words also do not contribute in the similarity computations which are used in most clustering methods (Aggarwal and Zhai, 2012). Subsequently, the software package CLUTO (http://glaros.dtc.umn.edu/gkhome/views/cluto) was used for automatic organization of the documents into clusters. As the clustering method, a CLUTO's implementation of the k-means algorithm was used together with

the H1 criterion function (Zhao and Karypis, 2001) and the cosine similarity measure. All these parameters were selected based on the previous experiments (Žižka, Burda and Dařena, 2012). As the number of clusters was not known in advance, different numbers between 20 and 1500 were tested. The cluster to which a document was assigned determined a document's topic.

Processing stock prices

Stock prices are represented by continuous numeric values constantly changing in time. However, especially for historical data, a few special values are important. They include, e.g., opening, closing, low, or high prices. Here, adjusted closing values that correct any distributions and corporate actions that occurred prior to the next day's open were used.

The absolute values are not of the same importance as the relative differences between certain moments in time. Therefore, a transformation based on stock price changes was applied. The prices change naturally very rapidly, usually in small proportions, reflecting many different events, habits, or sentiment (Blau and Griffith, 2016), or being completely random (Borch, 1963). Not all of the changes are therefore important as trends, cycles, or their combinations (Patel *et al.*, 2015). These significant movements might be thus revealed by replacing the original values by some values not showing that high volatility. Popular candidates are moving averages, often employed in technical analyses studying stocks markets. They substitute the original data by sequences of averages calculated from subsets of the data sets (a new value is calculated based on n last original values). Smoothing of the time series has proven to be a reasonable step. Moving averages based on a larger number of days (20 days) had more positive impact than moving averages based on a smaller number of days. The difference between the Simple Moving Average and the Exponentially Weighted Moving Average (NIST/SEMATECH, 2016) were negligible (Dařena *et al.*, 2018).

Not every change that can be easily measured between different moments is usually important. Wuthrich *et al.* (1998) found that appreciation and depreciation take place when the market moves up or down by at least 0.5 % and that the average change in market indices is often much more, about 1.5.

The relative difference in stock prices between two days is thus the basis for a movement type determination. For a movement to be considered significant, a minimal percent change needs to be specified. When being within the limit the price is considered to be constant. If the price increased more than, e.g., 1 %, an increase is detected, if there is a significant price drop, the price decrease is found. We considered 1 % minimal price difference to have sufficient amount of data. The analysis also considered a one-day lag between the publication of a text and a stock price movement as the association for this lag was found to be the strongest for this specific data (Dařena *et al.*, 2018).

Calculating the association

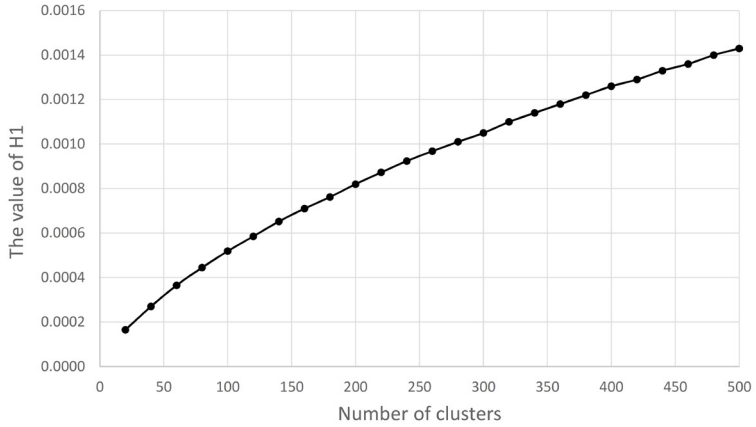
For each day and the investigated company, a stock price movement type derived according to the abovementioned procedure was known. In addition, the documents published a day ago and their topics were identified through the clustering process. Every day d for company c might be thus represented by a day-topics vector:

$$v_{c,d} = (n_{t1}, n_{t2}, \ldots, n_{tm}, m_{c,d}),$$

where n_{ti} is the number of documents with topic i related to day d and company c, and $m_{c,d}$ is a movement type of the stock price of company c in day d. Having enough such vectors for many companies and days, an association between documents' topics and stock price movements can be analyzed. If such an association exists, it should be describable by a function mapping the topics to stock price movement. This function is, however, not known and must be found. The process of finding such a function by generalizing many specific examples is known as induction and belongs to the family of supervised learning algorithms. Because the movement type is a discrete value, we can talk about classification. The process of mapping topics or their combinations to stock price movements is naturally not free of errors. The correctness of assigning movement types to topics, in other words, the correctness of classification, can be evaluated by usual classification performance measures that include accuracy, precision, recall, and F-measure (Sokolova *et al.*, 2006).

The data was massively unbalanced in terms of proportion of data items from individual classes. In a large majority of days, no significant change in stock prices was detected. In that case, biased or useless results in terms of accuracy would be achieved without further data set adjustment (Kubat *et al.*, 1998). Because significant movements of stock prices are more interesting, the interdependence between topics included in textual documents and stock price movements was investigated only in the periods with substantial price changes.

To describe the mapping from texts to stock price movements, three classifiers that achieved most promising results in our previous experiments (Dařena *et al.*, 2018) were used. They included the multinomial naïve Bayes classifier, support vector machines using sequential minimal optimization (Platt, 1998), and multinomial logistic regression with a ridge estimator (le Cessie and van Houwelingen, 1992), all implemented in Weka (Frank *et al.*, 2016). The correctness of classification was measured by the classification accuracy calculated during 10-fold cross-validation.



1: *The values of the H1 criterion function achieved for different numbers of clusters.*

Characterizing the topics

Using an unsupervised approach for topics separation cannot bring a perfect result. The quality of the output is usually lower that one might desire. The reason is that the algorithms do not have any prior knowledge of the data. The assignment of documents to clusters might be different than the one achieved by a human expert because only he or she has a clear objective and can use some additional, external information (Weiss *et al.*, 2010).

The quality of the clusters can also be evaluated according to the task for which the clusters serve as an input. Here, it is the classification task where the clusters (topics) play the role of attributes characterizing significant stock price movements. When the correctness of predictions using these attributes as independent variables achieves a sufficient level we might say that the clustering process brought acceptable outputs.

For a human, an insight to the data might be also important and interesting. To answer the question of what really caused the stock price increase or decrease, further investigation of the generated clusters is needed. A human expert can examine the documents in individual clusters and consider whether they belong to the same topic. However, in case of processing many documents, such a manual process would be infeasible. It is not necessary to look at all of the documents but to examine only some of them. There exist several approaches for choosing the representative documents (Gelbukh *et al.*, 2003). A user might be presented with a document which is an average (close the centroid of the cluster), the least typical (close the border of the cluster, near the remaining clusters), or the most typical document (located near the border of the cluster, far from the other clusters). Sometimes, one document does not have to be informative enough and more documents selected using the same criterion might be used.

When the documents are too long to read (in our case some documents contained even several thousands of words), only having some representative features describing them can significantly help. We suggest a procedure based on the assumption that the semantic content of the generated clusters is given by words (terms) that are significant for expressing the meanings. Certain important terms relate to a specific topic while other significant words to different ones (Žižka and Dařena, 2011a; Dařena and Žižka, 2011b). According to Ferraro and Wanner (2012), two main strategies for finding these significant terms, that can be used to assign a label to a cluster, can be identified. In the internal cluster labeling, the label of the cluster is based solely on the content of the cluster. It can be simply a list of words that appear with a sufficient frequency. In differential cluster labeling, the label is determined by contrasting the cluster with other clusters. When a candidate label depends on a cluster more than on the other clusters, it is considered a good label for that cluster. Several statistical measures, often used to select suitable sets of features, such as the Chi-square statistics, Mutual Information, or Information Gain might be applied. The outcome is a list of document elements that well characterize the cluster and can help in determining what the documents in the cluster are mostly talking about.

The Chi-square statistics, which proved to be effective in our previous work (Barák *et al.*, 2015), measures the independence of two events, here, an occurrence of a word and occurrence of the class (cluster). It measures, how an expected frequency of the events differ from the real (observed) frequency. It is a normalized value so it can be used to compare the terms in the same category (Manning *et al.*, 2009; Yang and Pedersen, 1997). The values of the Chi-square statistics can be calculated according to the following formula:

$$\chi^2 = \frac{N_{all}\left(A_i D_i - C_i B_i\right)^2}{\left(A_i + C_i\right)\left(B_i + D_i\right)\left(A_i + B_i\right)\left(C_i + D_i\right)},$$

where A_i = the number of the documents that contain the term t and also belong to category c_i,

B_i = the number of the documents that contain the term t but do not belong to category c_i, C_i = the number of the documents that do not contain the term t but belong to category c_i, i.e., $N_i - A_i$, D_i = the number of the documents that neither contain the term t nor belong to category c_i, i.e., $N_{all} - N_i - B_i$, N_i = the total number of the documents that belong to category c_i, N_{all} = the total number of all documents from the training data, and c_i – a class (i = *1..m*). When a term and a category are completely independent, the value of this measure is zero. The features most important with respect to a given class have thus the highest value.

The idea behind selecting the most important words characterizing clusters can be applied also to the selection of most important topics with respect to stock price movements so only the topics influencing the stock prices the most can be studied.

RESULTS

Initially, we had about 118,000 newspaper articles related to all companies together with their related stock price movements. From them, only 6,430 documents were associated with a significant stock price movement. These documents were clustered into different numbers of clusters. We studied the values of the H1 criterion function (the function to be optimized during the clustering process). Its values related to different numbers of clusters are depicted in Fig. 1. We could see that there is a clear "elbow" (a place from which the value of H1 does not grow so fast) between 100 and 200 of clusters. A higher number of clusters than does not give much better modeling of the data set (Morozkov *et al.*, 2012) so considering a much higher number was not reasonable. We studied clustering solutions having up to 200 clusters in detail and then 300, 400, and 500 clusters.

The numbers of clusters (topics in the document collection) were used as attributes for generating the vectors representing individual days with significant stock price changes. The data set was balanced so we had the same number of price increases and decreases. The data collection to be used for building a classifier now had almost 2,000 items.

Three classifiers were applied to the data and trained and tested using 10-fold cross-validation. The achieved classifier accuracies (the proportion of

2: *The accuracies achieved by three different classifiers for the day-topics data prepared using different numbers of document clusters.*

I: *The detailed accuracies achieved by three different classifiers for the day-topics data prepared using different numbers of document clusters.*

No. of Clusters	Naïve Bayes	Logistic Regression	SMO
20	0.617	0.630	0.614
40	0.656	0.664	0.654
60	0.656	0.676	0.658
80	0.682	0.690	0.677
100	0.678	0.698	0.677
120	0.703	0.709	0.692
140	0.704	0.706	0.685
160	0.714	0.720	0.705
180	0.717	0.725	0.699
200	0.723	0.714	0.697
300	0.720	0.721	0.700
400	0.742	0.726	0.712
500	0.744	0.741	0.727

correctly classified instances) can be found in Fig. 2 and Tab. I. We can see that with lower numbers of clusters the performance of all classifiers was quite similar. With a higher number of clusters, the naïve Bayes classifier dominated the others. With a higher number of clusters, the classification accuracy was still growing or was almost constant behind 500 clusters (this is not included in Fig. 2). On the other hand, the topics for a high number of clusters were represented by only a few documents so the generality of them was limited.

To characterize the topics influencing the stock price movements, the content of clusters most contributing to the movements were examined. To reveal the most influential topics, as well as the most influential content, the feature selection procedure using the values of the Chi-square statistics, was used, as described in the Data and Methods section. Because this is a quite demanding activity, especially when many clusters need to be inspected, we focused only on the clustering solutions having 100 and 200 clusters. To even increase the understandability for a human, 3-grams (sequences of three consecutive words) were considered informative features. When looking at a few 3-grams with the highest value of the Chi-square statistics, a human expert can often understand the prevailing meaning of the given group. Examples of such groups together with the derived meanings can be found in Tab. II.

DISCUSSION

To map topics to stock price movements, a supervised machine learning approach employing standard classification performance metrics was used. The achieved classification accuracy was increasing with the increasing number of clusters (i.e., the number of topics). The growth was more rapid for lower numbers of clusters. For higher numbers of clusters, the accuracy was growing only very slowly. The point where the change of the slope of the accuracy curve was the most significant lied the close to the point which was optimal in terms of selecting the right number of clusters using the elbow method.

Compared to the situation where the untransformed texts are used to map the articles' content to stock price movements, this approach enables treating groups of documents as new, derived data elements. These new entities makes the entire model simpler in terms of the number of features, which speeds up the training and prediction processes. After some loss of information (replacing a document consisting of hundreds of words by one topic), we can still demonstrate a clear correlation between the investigated data while having new possibilities of further analysis.

The feature selection methods were applied to the documents assigned to clusters representing their prevailing topic in order to characterize these topics. Some of these topics were rather independent on a specific company or industry (e.g., positive earnings) while the other represented something typical for an industry (e.g., mining) or region (e.g., Greek banking sector). Some topics also related to a particular event (e.g., dieselgate). Sometimes, a clearly defined topic could not be naturally found only when inspecting its most significant features. Here, a deeper analysis of the content of the documents would be necessary. It has to be also noted that the revealed prevailing topics are always closely related to the available data and time. It is possible that the same event can cause exactly the opposite effect sometime in the future.

It is generally not clear whether the stock price movements are reactions to the facts contained in texts or vice versa. In this paper, we studied

II: *Groups of the most significant features characterizing topics and derived prevailing meaning of them*

Significant Features (3-grams)	Prevailing Topic
prostate cancer drug /the French drugmaker /diabetes and cardiovascular /late stage pipeline	drugs, pharmacology
positive earnings ESP /an earnings beat /has an earnings	positive earnings
surface mining equipment /coal copper iron /iron ore oil /and underground mining /joy global inc	mining
Bank of Greece /European central bank /bank and eurobank /George Georgiopoulos editing	Greek banking sector
Western Digital corp /acquisition of EMC /on october Sandisk	computers
Delta Air Lines /American Airline group /available seat mile /passenger revenue per /per available seat	airlines
natural gas prices /billion cubic feet /gas prices are /five year average /natural gas storage	gas
Consol Energy inc /CNX coal resources /of million tons /Buchanan mine in /cubic feet equivalent	coal
Environmental protection agency /of Volkswagen s /chancellor Angela Merkel /of its diesel /s biggest carmaker /of the scandal	automotive, dieselgate
Under Armour ua /Nike and Under /apparel and footwear /college football championship/the sports apparel	sportswear

how financial markets react to the content of newspaper articles, which is a long-lasting question in finance (Wong, Liu and Chiang, 2014). Because the classification accuracy quantifying the strength of the relationship achieved an acceptable level it can be assumed that this assumption was appropriate. A similar research approach when the text and financial market data are shifted in the opposite direction would be also possible. In case of providing a proof of an existing relationship, it is possible that different topics (types of reactions to stock price movements) would be discovered. This was, however, not researched in this paper.

CONCLUSION

In our work, we focused on finding an interconnection between the content of the company related newspaper articles and the movements of stock prices of the respective companies. We needed to convert the original raw data to the formats suitable for further analysis. The stock prices were smoothed using a moving average and then transformed to one of three different types of a movement, given the minimal percentage change between two days. The documents were transformed, using an unsupervised clustering procedure, to their prevailing topics. For each day and company, we were thus able to generate a vector containing the information about what topics appeared in the news that day and what was the subsequent stock price movement.

These vectors were used to find a mapping from topics to a stock price movement across all companies. As the principal tool, machine learning based classification was employed. The trained classifiers served as functions that assigned a movement to a combination of topics. Standard classification performance measures were used to measure the correctness of this process. We achieved an accuracy of stock price movement predictions higher than 70%. A feature selection procedure was applied to the features characterizing the topics in order to facilitate the process of assigning a label to the topic by a human expert. It was demonstrated that the procedure generated meaningful topic related to the activities on stock markets.

Forthcoming research will focus on the automation of the process of characterizing the topics identified in the newspaper articles as the length and the number of the articles does not enable us to make a simple conclusion without an enormous effort. From the machine learning perspective, processing the data in a stream taking the time dimension into consideration, using, e.g., a moving window approach (Žižka and Dařena, 2015), and processing unbalanced data are attractive.

Acknowledgements

This research was supported by the Czech Science Foundation [grant No. 16-26353S "Sentiment and its Impact on Stock Markets"].

REFERENCES

AGGARWAL, C. C. and ZHAI, C. 2012. A survey of text clustering algorithms. In: AGGARWAL, C. C. and ZHAI, C. (Eds.). *Mining text data*. New York, NY: Springer, pp. 77–128.

BARÁK, K., DAŘENA, F. and ŽIŽKA, J. 2015. Automated Extraction of Typical Expressions Describing Product Features from Customer Reviews. *European journal of business science and technology*, 1(2): 83–92.

BENESTY, J., CHEN, J., HUANG, Y. and COHEN, I. 2009. *Pearson Correlation Coefficient*. Springer.

BLAU, B. M. and GRIFFITH, T. G. 2016. Price clustering and the stability of stock prices. *Journal of Business Research*, 69(10): 3933–3942.

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

BSOUL, Q., SALIM, J. and ZAKARIA, L. Q. 2013. An Intelligent Document Clustering Approach to Detect Crime Patterns. *Procedia Technology*, 11: 1181–1187.

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

LE CESSIE, S. and VAN HOUWELINGEN, J. C. 1992. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1): 191–201.

DAŘENA, F., PETROVSKÝ, J., ŽIŽKA, J. and PŘICHYSTAL, J. 2018. Machine Learning-Based Analysis of the Association between Online Texts and Stock Price Movements. *Inteligencia Artificial*, 21(61): 95–110.

DHILLON, I. S. and MODHA, D. S. 1999. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42: 143–175.

FERRANO, G. and WANNER, L. 2012. Labeling Semantically Motivated Clusters of Verbal Relations. *Procesamiento del Lenguaje Natural*, 49: 129–138.

FRANK, E., HALL, M. A. and WITTEN, I. H. 2016. *The WEKA Workbench*. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann.

GELBUKH, A. F., ALEXANDROV, M., BOUREK, A. and MAKAGONOV, P. 2003. Selection of Representative Documents for Clusters in a Document Collection. In: *Proceedings of Natural Language Processing and Information Systems*, *8th International Conference on Applications of Natural Language to Information Systems*, 120–126.

GUO, Q. and ZHANG, M. 2009. Multi-documents Automatic Abstracting based on text clustering and semantic analysis. *Knowledge-Based Systems*, 22(6): 482–485.

JOACHIMS, T. 2002. *Learning to classify text using support vector machines*. Norwell, MA: Kluwer Academic Publishers.

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

KUBAT, M., HOLTE, R. C. and MATWIN, S. 1998. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3): 195–215.

KUMAR, B. S. and RAVI, V. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114: 128–147.

LEE, H., SURDEANU, M., MACCARTNEY, B. and JURAFSKY, D. 2014. On the Importance of Text Analysis for Stock Price Prediction. In: *LREC*, pp. 1170–1175.

LI, X. et al. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69: 14–23.

LOUGHRAN, T. and MCDONALD, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66: 35–65.

MANNING, C. D., RAGHAVAN, P. and SCHÜTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

MOROZKOV, M., GRANICHIN, O., VOLKOVICH, Z. and ZHANG, X. 2012. Fast algorithm for finding true number of clusters. Applications to control systems. In: *Control and Decision Conference (CCDC)*, pp. 2001–2006.

NIST/SEMATECH. 2016. *e-Handbook of Statistical Methods*. [Online]. Available at http://www.itl.nist.gov/div898/handbook. [Accessed: 2016, August 11].

PATEL, J., SHAH, S., THAKKAR, P. and KOTECHA, K. 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1): 259–268.

PLATT, J. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: SCHOELKOPF, B., C. BURGES, C. and SMOLA, A. (Eds.). *Advances in Kernel Methods – Support Vector Learning*. MIT Press.

RANCO, G. et al. 2015. The effects of Twitter sentiment on stock price returns. *PloS one*, 10(9): e0138441.

SALTON, G. and MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

SCHUMAKER, R. P. and CHEN, H. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, 27(2): a12.

SIGANOS, A., VAGENAS-NANOS, E. and VERWIJMEREN, P. 2017. Divergence of sentiment and stock market trading. *Journal of Banking & Finance*, 78: 130–141.

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

TSENG, Y.-H., LIN, C.-J. and LIN, Y. 2007. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5): 1216–1247.

WEISS, S. M. et al. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.

WENG, B., AHMED, M. A. and MEGAHED, F. M. 2017. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, 79: 153–163.

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

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

YANG, Y. and PEDERSEN, J. O. 1997. A Comparative Study on Feature Selection in Text Categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420.

ZHAO, Y. and KARYPIS, G. 2001. *Criterion Functions for Document Clustering: Experiments and Analysis*. Technical Report #01-40. University of Minnesota, Department of Computer Science.

ZUO, Y. et al. 2016. Topic Modeling of Short Texts: A Pseudo-Document View. In: *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 2105–2114.

ŽIŽKA, J. and DAŘENA, F. 2011a. Mining Significant Words from Customer Opinions Written in Different Natural Languages. In: *Proceedings of the 14th International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*. Heidelberg: Springer, pp. 211–218.

ŽIŽKA, J. and DAŘENA, F. 2011b. Mining Textual Significant Expressions Reflecting Opinions in Natural Languages. In: *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications*, pp. 136–141.

ŽIŽKA, J., BURDA, K. and DAŘENA, F. 2012. Clustering a very large number of textual unstructured customers' reviews in English. In: *Proceedings of Artificial Intelligence: Methodology, Systems, and Applications*. Heidelberg: Springer, pp. 38–47.

ŽIŽKA, J. and DAŘENA, F. 2013. Revealing Prevailing Semantic Contents of Clusters Generated from Untagged Freely Written Text Documents in Natural Languages. In: *Text, Speech, and Dialogue*. Heidelberg: Springer, pp. 434–441.

ŽIŽKA, J. and DAŘENA, F. 2015. 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*. IEEE, pp. 131–138.

Contact information

František Dařena: frantisek.darena@mendelu.cz
Jan Přichystal: jan.prichystal@mendelu.cz