Examining Stock Price Movements on Prague Stock Exchange Using Text Classification

Jonáš Petrovský, Pavel Netolický and František Dařena
Department of Informatics, Faculty of Business and Economics, Mendel University in Brno,
Zemědělská 1, 613 00 Brno, Czech Republic
jontesek@gmail.com, pavel.netolicky@gmail.com, frantisek.darena@mendelu.cz

ABSTRACT

The goal of the article was to examine the relationship between the content of text documents published on the Internet and the direction of movement of stock prices on the Prague Stock Exchange. The relationship was modeled by text classification. As data were used news articles and discussion posts on Czech websites and the value of the PX stock index and stock price of company CEZ. Document's class (plus/minus/constant) was determined by the relative price change that happened between the publication date of a document and the next working day. We achieved a high accuracy of 75% for classification of discussion posts, however the classification accuracy for news articles was about 60%. We tried both binary (documents with constant class were discarded) and ternary classification – the former was in all cases more successful.

KEYWORDS

Text Mining, Classification, Stock Market, Machine Learning

1 INTRODUCTION

Efficient markets theory (EMT) says that investors immediately incorporate all available information about given stock into its price and therefore the stock price is based solely on its fundamental value. However empirical observations contradict the EMT, because some price movements cannot be explained by change of fundamental figures [1]. Here comes the Behavioral finance theory which says that emotions may deeply influence behavior and decision making of individuals as well as whole human societies [2]. This means that prices on capital markets are (more or less) influenced by emotions, moods and opinions of market participants [3]. These characteristics are difficult to obtain, but could be present in text documents published on the internet (news articles, social media posts, etc.), which express both fundamental facts (rationality) and emotions and opinions of people (irrationality) [4]. To determine if the texts actually contain such information and that it is connected with stock price, we need to show and quantify a connection between texts and stock price movements. In other words, examine the influence of the (i)rationality of investors on stock market.

2 CURRENTLY USED METHODS

When trying to model behavior of a stock price we can use classification or regression. Many studies (e.g. [4]) in this area chose the former approach and decided to examine not the actual numeric price value, but only change of the value – they used direction of the change (up, down or constant) as a class. We focused on this approach as well, because we are not interested in the actual stock price value, but only in its change.

In fact, the problem represents a typical text classification task – given a document, determine to which class it belongs. For this, virtually any supervised learning algorithm may be used. However there are two main difficulties. Firstly, documents' classes have to be defined in a meaningful and useful way. Lee et al. [5] used 1% threshold value for determining the direction of a stock price change. Secondly, a suitable and effective set of features must be chosen. Many studies used as features just single words (unigrams) in so-called bag-of-words model with satisfactory results [6]. The studies used different types of classifiers. Strength of the connection between texts and stock prices was evaluated by classification metrics (e.g. by accuracy) which are based on how many times the classifier assigns correct class to the given text.

3 DATA AND METHODOLOGY

The goal of the work was to examine the connection between content of text documents published on the Internet and direction of stock price movements, by using classification. A suitable approach had to be taken for working with every aspect of this task: handling prices and texts and processing the data via classification algorithms.

3.1 Stock prices

For the main part of the work, we used the PX index which reflects all companies traded on the Prague Stock Exchange (BCPP). The data were downloaded from the stock exchange’s website (https://www.pse.cz). For every trading day, we used the closing value of the index. We also decided to examine discussion posts for one company (CEZ). Because BCPP contains data only since 2012, we downloaded it from www.akcie.cz.

3.2 Text data

The examined text data (documents) were downloaded from two sources (see Table 1). All documents were written in Czech language.

Table 1. Examined text data.

Source	Documents type	Number of doc.	Period	Average per day
Patria.cz	News articles about Czech stock market.	1 244	9. 2. 16 to 27. 5. 17 (15 mon.)	2.63
Akcie.cz	Discussion posts about 17 Czech companies.	20 605	14. 3. 08 to 27. 5. 17 (9y.)	6.13

Table 2 shows the information available for every text document. In subsequent analysis, all these fields apart from author were used.

Table 2. Available characteristics of a document with a concrete example of a discussion post regarding company CETV.

Field name	original in Czech	translated to English
datetime	2017-05-18 11:49:00	2017-05-18 11:49:00
author	mmmm	mmmm
title	Za vodou koncila na 94.	Offshore price ended on 94.
text	ja si myslim ze se dostane nekam k 85 ale nemam kouli samozrejme. :))	I think that it will get to 85, but I don’t have crystal ball of course. :))

For every discussion post (Akcie.cz), it was known to which company on the stock exchange it belongs. However, for news articles (Patria.cz) this information was unknown. Moreover, it was found out, that a news article usually comments on multiple companies.

3.3 Classification methodology

We used classification to predict, whether a stock price will move up, down or stay constant on the basis of document’s text. Each price movement represented a class. To obtain more diverse and possibly better results, we used both two (only up and down) and three classes for classification. It was expected that the ternary classification would perform worse, like mentioned in [7]. We extracted documents' features from the text by using the bag-of-words model. Every document was represented by a vector with values corresponding to the assigned weights of the words present in the document. For the experiments with all discussion posts (Akcie.cz) and news articles (Patria) values of the PX index were used. For one experiment (referred to as “CEZ experiment”) stock prices and discussion posts related to only one company (CEZ) were used.

Document class. Assigning a class to a document was based on the relative price change between two moments and on the threshold value (v) of minimal percentage price change. Formally, the percentage price return R in time t is:

$$R_t = (p_t - p_{t-1}) / p_{t-1}, \qquad (1)$$

where price p_{t-1} is the closing price of the day when the document was published (or the last working day) and price p_t is close price the closing price of next working day. If the price return was in the constant interval (–v, +v), the document was either discarded from further processing (for binary classification) or assigned the “constant” class label (for ternary classification). If the price return was equal to or larger than +v, the document was labeled as “plus”, otherwise as “minus”. We used 0.25, 0.5 and 1.0% as the threshold values.

Text pre-processing and conversion. The text was processed as follows:
1. Join document title and text into one string.
2. Strip diacritics from text (convert “special” Czech letters to their ASCII equivalents).
3. Strip all HTML tags.
4. Lowercase and remove punctuation.
5. Tokenize – get words (using *TreebankWordTokenizer*).
6. Filter words – minimal length of 3 letters, exclude numbers.

The edited text had to be converted into a structured format. For this, a Python library called *scikit-learn* and its Vectorizer class were used. Only words which occurred at least 5 times (for discussion posts) and 10 times (for news articles) in the whole document collection were considered. Those words were converted to a bag-of-words representation using three different weighting schemes [8, p. 21–26]:

- Term Presence (TP): 1 if a term is present in a document, 0 if not.
- Term Frequency (TF): how many times is a term is present in a document.
- TF-IDF: TF (local weight) multiplied by IDF (global weight).

Classification. Converted data were processed again by *scikit-learn*. The data were split into training (60%) and testing (40%) datasets. Class balancing was not performed. Each of the generated vector representations was processed by 20 different classifiers (with default settings – we did not optimize the parameters of the classifiers). The performance of a classifier was rated by the achieved accuracy (proportion of correctly classified instances on all examined instances [9, p. 268]) on the test set.

4 RESULTS AND DISCUSSION

Three different sets of text data, all discussion posts (Akcie.cz), posts related to the CEZ company, and news articles (Patria) together with information about stock prices were used to prepare data for classification. Based on the combination of variable experimental parameters – the number of classes (2 or 3), minimal percentage change (0.25, 0.5 and 1%) and weighting scheme for the term-document matrix (TP, TF, TF-IDF) – 54 different sets were created and subsequently processed by 20 classification algorithms.

In total 1080 classification results were obtained. We evaluated the results for each data set separately and for each classification set, the highest accuracy achieved by any combination of vector type and classification algorithm was found. Our findings are presented in Table 3, Table 4 and Table 5. Class 1 means “minus”, class 2 “constant” and class 3 “plus”.

Table 3. Classification of Akcie.cz discussion posts

Num. of classes	Percent change	Accuracy	Total samples	Class 1 samples	Class 2 samples	Class 3 samples	Num. of words
2	0.25	0.74	16 673	8 169	0	8 504	10 604
2	0.50	0.76	13 449	6 454	0	6 995	9 181
2	1.00	0.78	8 170	3 939	0	4 231	6 359
3	0.25	0.67	20 576	8 169	3 903	8 504	12 337
3	0.50	0.65	20 576	6 454	7 127	6 995	12 337
3	1.00	0.79	20 576	3 939	12 406	4 231	12 337

Table 4. Classification of Patria news articles

Num. of classes	Percent change	Accuracy	Total samples	Class 1 samples	Class 2 samples	Class 3 samples	Num. of words
2	0.25	0.62	874	360	0	514	2 347
2	0.50	0.59	587	246	0	341	1 788
2	1.00	0.61	222	100	0	122	872
3	0.25	0.40	1 244	360	370	514	3 036
3	0.50	0.52	1 244	246	657	341	3 036
3	1.00	0.79	1 244	100	1 022	122	3 036

Table 5. Classification of CEZ discussion posts

Num. of classes	Percent change	Accuracy	Total samples	Class 1 samples	Class 2 samples	Class 3 samples	Num. of words
2	0.25	0.71	6 162	2 929	0	3 233	5 235
2	0.50	0.72	5 300	2 494	0	2 806	4 601
2	1.00	0.76	3 935	1 797	0	2 138	3 714
3	0.25	0.63	7 281	2 929	1 119	3 233	5 923
3	0.50	0.65	7 281	2 494	1 981	2 806	5 923
3	1.00	0.80	7 281	1 797	3 346	2 138	5 923

If we look at how balanced the datasets are, we can say that for 2 classes they are in all cases relatively well balanced. For 3 classes, there is a clear misbalance. This can be seen especially in Table 6 where 82% of samples belongs to the constant class.

If we compare the classification accuracy for different datasets, we see that for discussion posts it is far higher (+10%) than for news articles. Interesting is that the accuracy obtained by training a classifier on all discussion posts and the PX index (Table 3) is higher than when using only posts and prices for one company (Table 5).

The highest accuracy was achieved always for 3 classes and 1% change. If we consider only 0.25 and 0.50% changes, accuracy for 2 classes is always better than for 3 classes. Generally, it can be said that the higher the percentage change, the higher the accuracy. However, this does not hold for Patria news articles with 2 classes (Table 4), where is the highest accuracy achieved for 0.25% change.

Table 6. Comparison of avg. accuracies for vector type

Vector type	Akcie.cz	CEZ	Patria
TP	0.67	0.63	0.51
TF	0.66	0.63	0.51
TF-IDF	0.67	0.63	0.53

Table 6 tells us that for discussion posts the used vector type was not very important, however for news articles the highest accuracy was achieved by TF-IDF.

Table 7. Comparison of avg. accuracies for classification algorithms

Akcie.cz – discussion posts		Patria – news articles	
Algorithm	Avg. acc.	Algorithm	Avg. acc.
ExtraTrees	0.72	LogisticRegression	0.56
MLP	0.72	CalibratedClassifier	0.56
RandomForest	0.71	SVC	0.56
LogisticRegression	0.71	LogisticRegression	0.53
LinearSVC	0.71	RidgeClassifier	0.53

Table 7 shows classifiers with the best average performance across all experiments. For discussion posts “Extremely randomized trees” ensemble method was the best, closely followed by “Multi-layer Perceptron” neural network. However, out of the best five algorithms for news articles, only one (LogisticRegression) was successful also for the posts. This indicates that for each type of document different algorithms are suitable.

5 CONCLUSION

The goal of the article was to examine the relationship between the content of text documents published on the Internet and the direction of movement of stock prices on the Prague Stock Exchange. For this, text classification was used.

The connection was found as demonstrated by the achieved classification accuracy. When using binary classification (documents with constant class were discarded), we achieved an accuracy of 75-78% for discussion posts and about 60% for news articles. For ternary classification, the accuracy was lower (about 65% and 40-50%). However, for all datasets was the accuracy, when using the highest 1% threshold for minimal price change, 80 %.

During the work, we encountered several problems. The most notable one was a rather small amount of available data – especially the news articles.

It must be noted that the goal was to examine if there is a connection between texts and stock prices, not to achieve the highest possible accuracy for each classification algorithm. Because of this, we used only default settings (parameters’ values) for the algorithms. An optimization of these parameters might bring us a few percent higher accuracy.

There are many options for further research in this area: use clustering/topic models (e. g. LDA) to find document classes based on their content; use bigrams or tri-grams as features; take into account the importance (popularity) of the document, use more values for minimal price change and also other time interval (more or less than 1 day).

REFERENCES

[1] SHILLER, R. J. From efficient markets theory to behavioral finance. The Journal of Economic Perspectives. 2003, vol. 17, no. 1, p. 83–104.

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

[3] KAPLANSKI, G. and LEVY, H. Sentiment and stock prices: The case of aviation disasters. Journal of Financial Economics. 2010, vol. 95, no. 2, p. 174–201.

[4] ARIAS, M., ARRATIA, A. and XURIGUERA, R. Forecasting with Twitter Data. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, vol. 5, no. 1, p. 8.

[5] LEE, Heeyoung, et al. On the Importance of Text Analysis for Stock Price Prediction. In: LREC. 2014. p. 1170-1175

[6] Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems.

[7] DARENA, F., PETROVSKY, J., ZIZKA, J. and PRICHYSTAL, J. Analyzing the correlation between online texts and stock price movements at micro-level using machine learning," MENDELU Working Papers in Business and Economics 2016-67, Mendel University in Brno, Faculty of Business and Economics. Available at: https://ideas.repec.org/p/men/wpaper/67_2016.htm

[8] WEISS, S. M., INDURKHYA, N. and ZHANG, T. Fundamentals of Predictive Text Mining. London: Springer, 2010.ISBN 978-1-84996-225-4.

[9] MANNING, C. D. and SCHÜTZE, H. Foundations of Statistical Natural Language Processing. Cambridge, USA: The MIT Press, 1999. ISBN 9780262133609.