Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction

Petr Hajek
Faculty of Economics and Administration
University of Pardubice
Studentska 84, 53210 Pardubice
Czech Republic
+420 466 036 147
petr.hajek@upce.cz

Aliaksandr Barushka
Faculty of Economics and Administration
University of Pardubice
Studentska 84, 53210 Pardubice
Czech Republic
+420 466 036 147
st47591@student.upce.cz

ABSTRACT
Media-expressed information in financial news are critical for stock market prediction. Nevertheless, researchers have primarily focused on the role of sentiment analysis in predicting stock returns and volatility. Here we show that topics discussed in the financial news may carry additional important information. We use a combination of sentiment analysis (using finance-specific dictionary-based approach) and topic detection (using latent dirichlet allocation) to predict one-day-ahead stock movements of major US companies. The proposed system employs a deep neural network to model complex stock market relations. We demonstrate the effectiveness of this approach compared to baselines, such as support vector machines and sentiment- and topic-based models used separately.

CCS Concepts
• Information systems → Information systems applications → Decision support systems → Data analytics • Computing methodologies → Artificial intelligence → Natural language processing→Information extraction.

Keywords
Sentiment analysis; topic detection; financial news; stock movement.

1. INTRODUCTION
Textual analysis is an increasingly important area in financial prediction models. The past decade has seen its rapid development in various financial market problems, including stock return prediction [1-3] or volatility modelling [4,5]. These studies have empirically confirmed the importance of information hidden in textual documents for developing accurate prediction models. Textual documents such as annual reports, financial news or social media messages carry complementary have increasingly been used as the sources of information about the company's future financial development. It is assumed that this information manifests in stock market expectations and thus also in the development of stock market price. Automated systems that can analyze the company-related texts are therefore of critical importance. These systems are usually based on sentiment analysis of the texts. This can be conducted using either dictionary-based approach [4] or machine learning [5].

More recently, topic detection has also attracted attention in related literature because topics discussed in the financial news or social media may carry additional important information for investors [6]. However, the advantages of both textual analysis approaches have been studied only for corporate annual reports [7] and social media, namely message board [6]. These studies have shown that combining sentiment analysis and topic detection improves the accuracy of financial prediction models. We hypothesize that similar effect can be observed for financial news. Therefore, here we propose a system integrating sentiment analysis and topic detection. In conformity with the state-of-the-art literature [7], we employ a deep neural network to train this system to predict stock movements. We demonstrate that this approach is more effective not only than traditional support vector machine (SVM) method but also than sentiment analysis and topic detection used separately.

To conduct sentiment analysis, we adopt the approach of most previous studies and use predefined finance-specific dictionaries. More precisely, we calculate the sentiment score as the difference between positive and negative words. For topic detection, we employ latent dirichlet allocation (LDA), a generative probabilistic model that is used to detect a mixture of hidden topics in texts [6]. To perform the one-day-ahead prediction of stock movements (up or down), we use a deep feedforward neural network with dropout regularization and rectified linear units [7], hereinafter referred to as the DNN. This method overcomes the limitations of traditional machine learning methods used for the stock movement prediction, such as problem with overfitting and optimization convergence to a poor local minimum.

The remainder of this paper is organized in the following way. Section 2 reviews relevant literature on the relationship between textual analysis of financial news and stock market prediction. Section 3 presents our research methodology, including the data and its pre-processing. Section 4 presents the results of experiments and section 5 concludes the paper.

2. RELATED LITERATURE
Financial news represent an important source of media-expressed information. The effect of financial news' content on future stock returns has received much attention in last decade [1-3]. More precisely, sentiment analysis of financial news has become a critical issue in the literature related to stock markets. Researchers have primarily focused on positive and negative sentiment in the news and its short-term effect on future stock returns [8,9]. This effect has profound theoretical background in behavioral finance, namely in the prospect theory [10], which emphasizes that a firm's stakeholders process information asymmetrically, suggesting the critical role of emotional information they process.

Financial news are relevant to both the overall economic and financial market conditions and to individual firms [1]. The former are used to analyze the behavior of the whole stock markets, while the latter are suitable for predicting individual stock prices, returns, volatilities and other firm-level indicators.

Tetlock et al. [8] examined the effect of financial news on future stock returns, showing that negative sentiment in firm-related

financial news are effective in predicting low stock returns. The negative effect of pessimism in financial news was also reported by [9]. Similar findings were observed for intra-day trading [11]. Positive and negative sentiment in financial news was reported to be valuable for short sellers who are effective in processing information [12]. In addition to sentiment, subjective tone of the news was found to be important when predicting stock price direction [13]. Several trading strategies based on sentiment analysis of financial news were tested by [14] in order to maximize profits of automated investment systems. Kelly and Ahmad [15] incorporated domain-specific news sentiment into a simple buy and hold trading strategy to increase annual returns.

A wider textual representation was used by [16], including Bag-of-Words (BoW), named entities and noun phrases. When combined with stock prices, the prediction model performed best with these representations. Market feedback was used for text features' selection in [17]. This process significantly improved the prediction performance of the used support vector machine (SVM) model. Higher trading return was obtained by using SVM model based on both financial news and social media content [18]. In a similar model, it was shown that c characteristics also affect the SVM prediction accuracy [19]. Li et al. [20] performed a comparative analysis of BoW-based SVM model and rule-based sentiment model (based on dictionaries/word lists), demonstrating that the latter one performs better. Alternatively to the BoW approach, DNNs were applied to sentiment analysis of financial news by [21], thus improving the predicting accuracy of stock price movements. A machine learning model using multilayer perceptron (MLP) neural network was employed by [22] to show that the model performed best when sentiment scores were combined with categorization of the news into business events. Topic detection also exhibited large effects on abnormal stock return prediction in [23]. Specifically, forty topics were found in the corpus of financial news using LDA method [23]. For a comprehensive overview of the methods used for sentiment analysis and topic detection in financial domain, see [24].

3. DATA AND RESEARCH METHODOLOGY

The framework of research methodology used in this study is depicted in Figure 1.

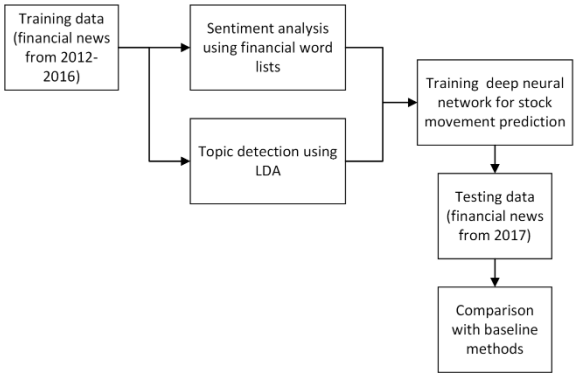

Figure 1. Research methodology.

First, financial news for each company were downloaded from the Reuters database (www.reuters.com) for the period of six years (2012-2017). The data for the period 2012-2016 were used as training data, while those for the year 2017 served as testing data. Specifically, we collected the data for 15 US companies listed at the New York Stock Exchange with the largest market capitalizations (the ticker symbols of these companies are presented in Table 1). When multiple financial news were available at the date of disclosure, the news were merged into one document. In case of missing news, the corresponding transaction was removed from the dataset. Also note that prediction models were trained for each company separately to consider company characteristics [19].

As finance-specific word lists performed better compared to general dictionaries in previous studies [20], here we used the word lists proposed by Loughran and McDonald [4]. These include 354 positive and 2,329 negative words (available at https://sraf.nd.edu/textual-analysis/resources/). In agreement with previous studies [7,22,25], we calculated the raw frequencies of positive (POS) and negative (NEG) words in the company-related financial news. The problem of negations of positive words was addressed by using collocation analysis with negation words in QDA Miner 5. After subtracting these from the number of positive words, we calculated the sentiment score as follows:

$$(POS - NEG) / (POS + NEG). \quad (1)$$

The average sentiment score was 0.674, indicating that positive sentiment prevailed in the financial news. Detailed results for each company are reported in Table 1.

To perform the LDA, we used the Stanford Topic Modeling Toolbox 0.4.0. Standard pre-processing techniques were used, including the removal of stop words and lemmatization. To control for feature selection bias, the LDA was carried out on the training data first. Then, the topics were inferred for testing data by using Gibbs sampling. To select the number of topics, we evaluated the perplexity (the average number of equiprobable word choices on unseen data). A rule of thumb was used to choose the number of topics for which the perplexity started to decrease. On average, we detected 24.6 topics per company.

The movement of each stock was calculated from daily close prices. Specifically, if daily close price at day t+1 was higher than that at day t, the transaction was categorized as +1 (up), otherwise as -1 (down). Historical data from Yahoo! Finance (finance.yahoo.com) were used to calculate the stock movements. Note that the textual analyses were performed for the financial news obtained at day t.

Table 1. Average sentiment scores and numbers of topics

Stock	Aver. sentiment	No. of topics	Stock	Aver. sentiment	No. of topics
BAC	0.63	23	T	0.70	24
CVX	0.64	21	PG	0.71	38
XOM	0.63	22	VZ	0.72	24
WFC	0.74	24	KO	0.68	26
WMT	0.62	18	HD	0.71	38
PFE	0.66	21	GE	0.73	20
JNJ	0.64	26	DIS	0.69	23
JPM	0.61	21			

To predict the stock movements, we employed the DNN model defined as follows. The model is represented by the multilayer perceptron with multiple hidden layers, which are used to model complex relations between the input attributes and output classes. To avoid overfitting of this complex DNN structure, dropout

regularization was used. This is, neurons (units) were randomly dropped from the DNN to address this problem. Another problem related to traditional neural network models is how to avoid poor local minima and slow convergence of error function. To overcome this problem, rectified linear units were used instead of sigmoidal units, see [26] for details. To train the DNN, we employed the mini-batch gradient descent algorithm defined as follows:

$$w_{k+1} = w_k - \eta \nabla_\theta J(w_k; x^{(i:i+n)}; y^{(i:i+n)}), \quad (2)$$

where w is synapse weight, k denotes iteration, η represents learning rate, J denotes an objective function, x^i are inputs, and y^i is the output of the data sample i in the mini-batch, i=1, 2, …, n. The mini-batches ensure stable convergence of the learning process.

As the baseline method, we employed SVM, a predominantly used machine learning method in previous studies [17-19]. Furthermore, we compared the results of the DNN with those obtained without sentiment analysis and topic detection, respectively. Thus, the performance can be compared with the previous research based on either of these textual analysis methods. In addition, such comparison demonstrates the sensitivity of the results to sentiment and topic information, respectively. To measure the quality of the prediction, we used a standard classification measure for imbalanced datasets, namely AUC (area under receiver operating characteristic curve). In fact, the ratio of up movements ranged from 39 to 57 % for individual stocks. Therefore, alternative measures like accuracy may yield misleading conclusions.

4. EXPERIMENTAL RESULTS

To achieve high prediction accuracy in terms of AUC, we carried out many experiments using DNN and SVM under different settings in software Weka 3.8. Specifically, we optimized both the structure of the DNN model and its learning process using grid search procedure as follows. Regarding the DNN structure, the tests were carried out for different numbers of hidden layers {1, 2} and numbers of units in the hidden layers {5, 10, 20, 50}. For effective dropout process, we used the dropout rate of 0.2 for input layer and 0.5 for hidden layers [7]. Similarly, default setting of the learning process included the learning rate of 0.05, 1000 iterations and the size of mini-batches of 100. The SVM used in this study for the purpose of comparative analysis was trained using stochastic gradient descent (SGD) algorithm. This method has several striking advantages, such as robustness to data sparsity and high dimensionality. The SGD was tested for 500 iterations and the learning rate of 0.05.

Table 2 shows the results obtained on testing data (year 2017) for several settings. Both the DNN and SVM were tested using different input attributes: (1) sentiment score only (DNN-S and SVM-S), (2) topic values (DNN-T and SVM-T), and (3) combination of sentiment score and topic values (DNN-S+T and SVM-S+T). The experiments were performed for each of the fifteen stocks. The overall average performance suggests that all methods perform better than random classification (AUC>0.500). The best average performance was achieved by using DNN-S+T, followed by the topic model DNN-T and sentiment model DNN-S, respectively. Although the SVM models were outperformed by those of the DNN, the combined sentiment and topic SVM model was followed by SVM-T and SVM-S. As expected, the integrated models performed better than those trained using either sentiment or topic information. Although more information seems to be carried by the models based on topic detection, sentiment score provided additional value. Moreover, when focusing on individual stocks, sentiment-based models performed best in case of PFE, VZ and HD stocks. Similarly, topic-based models were most effective for predicting movements of several stocks (JNJ, JPM and GE). For the remaining stocks, the combination of sentiment and topic models achieved highest AUC. Furthermore, the DNN models outperformed SVM in 12 out of 15 stocks. In fact, for most stocks SVM was not effective in learning sentiment-based models at all. More precisely, AUC=0.500 suggest that all data samples were classified into the majority class only. Overall, our assumption concerning the effect of company characteristics was confirmed as the evaluated methods performed differently across companies.

Table 2. Comparison of methods in terms of AUC

Stock	SVM-S+T	SVM-S	SVM-T	DNN-S+T	DNN-S	DNN-T
BAC	0.487	0.510	0.517	*0.595*	0.485	0.529
CVX	0.579	0.557	0.556	*0.644*	0.580	0.533
XOM	0.490	0.500	0.496	*0.526*	0.483	0.510
WFC	0.510	0.500	0.504	*0.537*	0.493	0.510
WMT	0.467	0.500	0.489	*0.577*	0.514	0.515
PFE	0.475	*0.529*	0.475	0.509	0.523	0.494
JNJ	0.555	0.500	0.554	0.570	0.461	*0.608*
JPM	0.524	0.500	0.530	0.560	0.539	*0.589*
T	*0.578*	0.500	0.558	0.546	0.497	0.503
PG	0.611	0.500	0.626	*0.680*	0.562	0.593
VZ	0.481	0.523	0.504	0.533	*0.551*	0.528
KO	*0.621*	0.575	0.529	0.603	0.607	0.626
HD	0.479	0.531	0.406	0.542	*0.596*	0.464
GE	0.527	0.521	0.565	0.567	0.526	*0.596*
DIS	0.482	0.500	0.512	*0.592*	0.540	0.529
Mean	0.524	0.516	0.521	*0.572*	0.530	0.542
St.Dev.	0.050	0.023	0.048	0.044	0.042	0.047

In the next step, we compared the performance of the tested models statistically. Most of the AUC distributions were not normal so non-parametric Wilcoxon signed rank tests were run (Table 3). These tests show that the DNN-S+T significantly outperformed all the remaining models at p=0.05. SVM models were further significantly outperformed by the DNN-T, which performed similarly as the DNN-S. Finally, all the SVM models performed statistically similar in terms of AUC.

In Table 4, we present the performance of the DNN-S+T model in terms of average return achieved in the year 2017. This return is calculated using simple strategy, with buy order invoked by +1 prediction and sell order by -1 prediction. For the comparative purposes, we also calculated the performance of traditional buy and hold strategy (B&H). For the sake of simplicity, fees were not taken into account in these calculations. The DNN-based investment strategy not only achieved more than four times higher average return than the B&H strategy but it was also less risky with about half the standard deviation of the return. On the one hand, the DNN-based strategy was not effective for several stocks (XOM, PFE, JNJ and T). What was common to all these stocks was that the DNN

predominantly predicted the -1 class (down). In other words, this strategy was too cautious in several cases. On the other hand, the strategy achieved non-negative return for all stocks except for the GE.

Table 3. Wilcoxon signed-rank test for AUC

Method	SVM -S+T	SVM -S	SVM -T	DNN -S+T	DNN -S	DNN -T
SVM-S+T		0.647	1.000	1.000	1.000	1.000
SVM-S	1.000		1.000	1.000	1.000	1.000
SVM-T	0.601	0.584		1.000	1.000	1.000
DNN-S+T	0.002#	0.001#	0.001#		0.011#	0.044#
DNN-S	0.864	0.118	0.842	1.000		1.000
DNN-T	0.088*	0.055*	0.065*	1.000	0.293	

significant at p=0.05, * at p=0.1.

Table 4. Return [%] of B&H and DNN-based investment strategies

Stock	Return B&H	Return DNN	Stock	Return B&H	Return DNN
BAC	12.26	20.00	PG	-0.89	0.83
CVX	-3.64	8.62	VZ	-7.66	0.00
XOM	7.23	0.68	KO	2.33	3.26
WFC	-0.13	4.81	HD	1.13	1.13
WMT	19.03	24.67	GE	-47.88	-7.51
PFE	8.76	1.67	DIS	-7.14	2.88
JNJ	12.76	5.73	Mean	1.24	5.76
JPM	14.13	14.23	St.Dev.	15.22	7.97
T	8.31	5.32			

5. CONCLUSIONS

In conclusion, the evidence from this study suggests that integrating sentiment analysis and topic detection in financial news improves the performance of stock movement prediction models. In this investigation, the aim was also to assess the performance of DNN in this prediction problem. This study has found that generally DNN outperforms tradition SVM models. However, the findings to emerge of this study is that the results largely depend on the company characteristics. More precisely, the balance of the stock movement classes seems to be of critical importance.

The empirical findings in this study add to a growing body of literature on the role of textual analysis in stock market modelling. Specifically, our findings suggest a role of sentiment and topic information hidden in financial news for stock movement prediction. However, a number of caveats need to be noted regarding the present study. This study was unable to analyse the effect of technical indicators in the prediction model. Further research might explore their effect in a more complex prediction model. More qualitative information can also be obtained by using additional textual sources such as social media. Different prediction horizons can be investigated in a future study. Finally, our results should be validated by a larger sample size of companies.

6. ACKNOWLEDGMENTS

This study was funded by the scientific research project of the Czech Sciences Foundation (grant number GA16-19590S).

7. REFERENCES

[1] Kearney, C. and Liu, S. 2014. Textual sentiment in finance: A survey of methods and models. *Int. Rev. Financ. Anal.* 33, 171-185. DOI=10.1016/j.irfa.2014.02.006.

[2] Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. 2014. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* 41, 7653-7670. DOI=10.1016/j.eswa.2014.06.009.

[3] Loughran, T. and McDonald, B. 2016. Textual analysis in accounting and finance: A survey. *J. Account. Res.* 54, 1187-1230. DOI=10.1111/1475-679X.12123.

[4] Loughran, T. and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66, 35-65. DOI=10.1111/j.1540-6261.2010.01625.x.

[5] Myskova, R., Hajek, P., and Olej, V. (2018). Predicting abnormal stock return volatility using textual analysis of news - A meta-learning approach. *Amfiteatru Econ.* 20(47), 185-201.

[6] Nguyen, T. H., Shirai, K., and Velcin, J. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* 42(24), 9603-9611. DOI= 10.1016/j.eswa.2015.07.052.

[7] Hajek, P. 2018. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Comp. Appl.* 29(7), 343–358. DOI= 10.1007/s00521-017-3194-2.

[8] Tetlock, P. C., Saar-Tsechansky M., and MacSkassy, S. 2008. More than words: Quantifying language to measure firms' fundamentals. *J. Finance* 63, 1437-1467. DOI=10.1111/j.1540-6261.2008.01362.x.

[9] Tetlock, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *J. Finance* 62, 1139-1168. DOI=10.1111/j.1540-6261.2007.01232.x.

[10] Kahneman, D. and Tversky, A. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making*, Part I, 99-127.

[11] Li, X., Huang, X., Deng, X., and Zhu, S. 2014. Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. *Neurocomputing* 142, 228-238. DOI= 10.1016/j.neucom. 2014.04.043.

[12] Engelberg, J. E., Reed. A. V., and Ringgenberg, M. C. 2012. How are shorts informed?. Short sellers, news, and information processing. *J. Financ. Econ.* 105, 260-278. DOI=10.1016/j.jfineco.2012.03.001.

[13] Schumaker, R. P., Zhang Y., Huang C. N., and Chen, H. 2012. Evaluating sentiment in financial news articles. *Decis. Support Syst.* 53, 458-464. DOI=10.1016/j.dss.2012.03.001.

[14] Feuerriegel, S. and Prendinger, H. 2016. News-based trading strategies. *Decis. Support Syst.* 90, 65-74. DOI=10.1016/ j.dss.2016.06.020.

[15] Kelly, S. and Ahmad, K. 2018. Estimating the impact of domain-specific news sentiment on financial assets. *Knowl.-*

Based Syst. 150, 116-126. DOI=10.1016/j.knosys.2018.03.004.

[16] Schumaker, R. P. and Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news. *ACM Trans. Inf. Syst.* 27, 1-19. DOI=10.1145/1462198.1462204.

[17] Hagenau, M., Liebmann, M., and Neumann, D. 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decis. Support Syst.* 55, 685-697. DOI=10.1016/j.dss.2013.02.006.

[18] Li, Q., Wang, T., Gong, Q., et al. 2014. Media-aware quantitative trading based on public Web information. *Decis. Support Syst.* 61, 93-105. DOI=10.1016/j.dss.2014.01.013.

[19] Li, Q., Wang, T., Li, P., et al. 2014. The effect of news and public mood on stock movements. *Inf. Sci.* (Ny) 278, 826-840. DOI=10.1016/j.ins.2014.03.096.

[20] Li, X., Xie, H., Chen, L., et al. 2014. News impact on stock price return via sentiment analysis. *Knowl.-Based Syst.* 69, 14-23. DOI=10.1016/j.knosys.2014.04.022.

[21] Kraus, M. and Feuerriegel, S. 2017. Decision support from financial disclosures with deep neural networks and transfer learning. *Decis. Support Syst.* 104, 38-48. DOI=10.1016/j.dss.2017.10.001.

[22] Geva, T. and Zahavi, J. 2014. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decis. Support Syst.* 57, 212-223. DOI=10.1016/j.dss.2013.09.013.

[23] Feuerriegel, S. and Ratku, A. 2016. Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In Bui TX, Sprague RH (eds) *49th Hawaii Int. Conf. Syst. Sci. IEEE*, Kauai, 1072-1081.

[24] Kumar, B. S. and Ravi, V. 2016. A survey of the applications of text mining in financial domain. *Knowl.-Based Syst.* 114, 128-147. DOI= 10.1016/j.knosys.2016.10.003.

[25] Hajek, P. and Henriques, R. 2017. Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. *Knowl.-Based Syst.* 128, 139-152. DOI=10.1016/j.knosys.2017.05.001.

[26] Maas, A. L., Hannun, A. Y., and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML* 30(1), 1-6.

Authors' background

Your Name	Title*	Research Field	Personal website
Petr Hajek	associate professor	knowledge-based systems, financial modelling	https://fes.upce.cz/user/5408/F97730C4-344C-4A64-8EAB-99D8412787A7
Aliaksandr Barushka	Phd candidate	machine learning, text mining	https://fes.upce.cz/kontakty-usii