A sentiment analysis-based machine learning approach for financial market prediction via news disclosures

Raymond Chiong
The University of Newcastle
Callaghan, NSW 2308, Australia
Raymond.Chiong@newcastle.edu.au

Zongwen Fan
The University of Newcastle
Callaghan, NSW 2308, Australia
Zongwen.Fan@uon.edu.au

Zhongyi Hu
Wuhan University
Wuhan 430072, China
Zhongyi.Hu@whu.edu.cn

Marc T.P. Adam
The University of Newcastle
Callaghan, NSW 2308, Australia
Marc.Adam@newcastle.edu.au

Bernhard Lutz
University of Freiburg
Freiburg 79085, Germany
Bernhard.Lutz@is.uni-freiburg.de

Dirk Neumann
University of Freiburg
Freiburg 79085, Germany
Dirk.Neumann@is.uni-freiburg.de

ABSTRACT

Stock market prediction plays an important role in financial decision-making for investors. Many of them rely on news disclosures to make their decisions in buying or selling stocks. However, accurate modelling of stock market trends via news disclosures is a challenging task, considering the complexity and ambiguity of natural languages used. Unlike previous work along this line of research, which typically applies bag-of-words to extract tens of thousands of features to build a prediction model, we propose a sentiment analysis-based approach for financial market prediction using news disclosures. Specifically, sentiment analysis is carried out in the pre-processing phase to extract sentiment-related features from financial news. Historical stock market data from the perspective of time series analysis is also included as an input feature. With the extracted features, we use a support vector machine (SVM) to build the prediction model, with its parameters optimised through particle swarm optimisation (PSO). Experimental results show that our proposed SVM and PSO-based model is able to obtain better results than a deep learning model in terms of time and accuracy. The results presented here are to date the best in the literature based on the financial news dataset tested. This excellent performance is attributed to the sentiment analysis done during the pre-processing stage, as it reduces the feature dimensions significantly.

KEYWORDS

Financial market prediction, Sentiment analysis, Support vector machine, Particle swarm optimisation

ACM Reference Format:
Raymond Chiong, Zongwen Fan, Zhongyi Hu, Marc T.P. Adam, Bernhard Lutz, and Dirk Neumann. 2018. A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of Genetic and Evolutionary Computation Conference Companion (GECCO'18 Companion),* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article pos145, 2 pages. https://doi.org/10.1145/3205651.3205682

GECCO'18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07…$15.00
https://doi.org/10.1145/3205651.3205682

1 INTRODUCTION

Financial market prediction based on news disclosures is attracting more and more attention in recent years, because financial news has great influence in the stock market [3]. Investors often rely on financial news information to make their decision of buying or selling [2]. However, it is very difficult to predict the financial market accurately based on news disclosures, due to the complexity and ambiguity of natural languages used [6]. Recently, complex models such as deep learning and transfer learning were applied to predict stock market movements by Kraus and Feuerriegel using financial disclosures [5]. In their study, tens of thousands of features were extracted by using bag-of-words and term frequency-inverse document frequency (tf-idf) techniques, a very complex and time-consuming way to build a prediction model. With such a high-dimensional feature space, a deep learning model having more than 500,000 parameters can only achieve 57.8% of prediction accuracy.

In this study, to improve the efficiency of modelling, we propose the use of sentiment analysis to extract the most informative features. The idea of time series is also utilised, considering the chronological order of news disclosures. As a result, only a small number of input features are required to build our prediction model. We opt for a simple support vector machine (SVM) as the prediction model, given its strong capability of solving highly nonlinear problems [7] and wide applicability in financial market prediction [4]. We further use particle swarm optimisation (PSO) [1] to tune the parameters of this SVM-based forecasting model.

2 DATA PRE-PROCESSING AND SETUP

The dataset used in this study contains 13,135 regulated German ad hoc announcements in English [5]. To be consistent with the work of Kraus and Feuerriegel [5], which considered the chronological order of news disclosures, the first 80% of the timeframe was used as the training set, and the rest as the testing set.

In the data pre-processing stage, disclosures of penny stocks and those published on non-trading days were first removed. Sentiment analysis based on natural language processing was applied to analyse stock companies' disclosures, denoted as *Messages*. By

using the *TextBlob* library in Python, two features – *polarity* and *subjectivity* – were extracted from each *Message* disclosure. The value of *polarity* is between -1 and 1, while the value of *subjectivity* is between 0 and 1. Let *AbnormalReturn* denote the abnormal return computed based on a market model [5]. *AbnormalReturnSign* is the sign of abnormal return, which can be 1 (positive) or 0 (negative). For time series analysis, the 1-day lagged *AbnormalReturnSign* was used as an input feature. Counting the two sentiment features (*polarity* and *subjectivity*) and the lagged *AbnormalReturnSign*, we have three input features in total. With these input features, we aim to predict the *AbnormalReturn*. It is worth noting that our feature extraction procedure, which reduces the dimensionality of the dataset significantly, is different from that used by Kraus and Feuerriegel [5]. In our SVM-based prediction model, the stock market trend is encoded as 0 if the predicted *AbnormalReturn* is less than zero, or 1 otherwise.

3 SVM-BASED PREDICTION

Fig. 1 shows a flowchart of our SVM-based prediction model with sentiment analysis pre-processing and PSO parameter tuning. The kernel function for the SVM is a Gaussian kernel, determined by trial-and-error. Its three hyper-parameters, the penalty parameter of error term, bandwidth of Gaussian kernel, and epsilon-tube, were optimised through PSO. The swarm size and maximum iteration of PSO were set to 5 and 100, respectively.

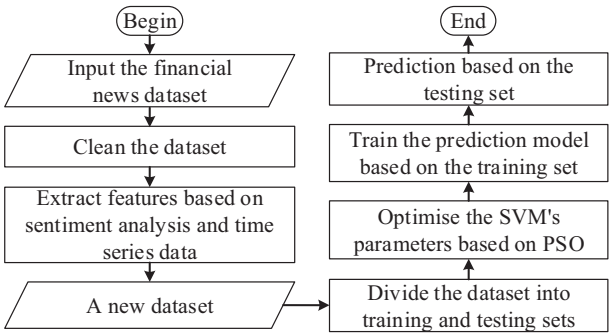

Figure 1: A flowchart of our prediction model.

4 RESULTS AND DISCUSSION

Experimental results in terms of accuracy and computational time are shown in Table 1. Kraus and Feuerriegel's deep learning model, denoted as DL_tfidf, is expectedly the most time-consuming. An SVM-based model with features also extracted by tf-idf, SVM_tfidf, requires less time than deep learning. By using features extracted via our sentiment analysis based approach, SVM_senti needs only 17s for the prediction task. The time taken by SVM_senti_GS and SVM_senti_PSO, which are SVM_senti-based forecasting models with hyper-parameters optimised by Grid Search and PSO, respectively, is still much less than the models using tf-idf pre-processing (i.e., SVM_tfidf and DL_tfidf). This is because the number of features we extracted (three features in total) is far less than tf-idf (tens of thousands of features).

Table 1: Experimental results based on accuracy and time.

Algorithm	Accuracy	Time (s)
DL_tfidf [5]	0.578	79200
SVM_tfidf	0.5452	15286
SVM_senti	0.5787	17
SVM_senti_GS	0.5856	1098 (64 settings)
SVM_senti_PSO	0.5915	8415 (100 epochs)

As for the accuracy, each reported based on the average of 20 runs, we see that SVM_senti is able to produce a higher accuracy value of 57.87% compared to SVM_tfidf's 54.52%. The reasons being 1) through sentiment analysis we are able to extract the most informative features from the financial news dataset, thereby reducing the effect of noise or less useful features substantially; and 2) the use of time-series pre-processing takes full advantage of historical information, which to some extent improves the accuracy. Based on our pre-processing approach, the accuracy of SVM_senti is also slightly better than the deep learning model by Kraus and Feuerriegel [5]. Using Grid Search or PSO for parameter optimisation enables us to outperform the deep learning model by a bigger margin (0.76% and 1.35%, respectively).

5 CONCLUSION

This paper presented a sentiment analysis-based machine learning approach for financial market prediction via news disclosures. Results showed that sentiment analysis is a good way to extract useful features from financial news data, which reduces the feature dimensions significantly. Using PSO for parameter optimisation further improves the prediction accuracy. The accuracy of 59.15% obtained by our SVM and PSO-based prediction model replaces the previous best result of 57.8% obtained by Kraus and Feuerriegel's deep learning model. Although we are unable to ascertain if differences between the results are statistically significant, as we have no access to Kraus and Feuerriegel's experimental data, the superior computational time of our proposed model is a clear indication that we can use a simpler approach to achieve comparable performance.

ACKNOWLEDGEMENTS

This work was supported through the Australia-Germany Joint Research Cooperation Scheme (G1600912) and Fundamental Research Funds for the Central Universities (Grant No. 104-413000017).

REFERENCES

[1] Maurice Clerc. 2010. *Particle swarm optimization.* Vol. 93. John Wiley & Sons.
[2] Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL*, Vol. 7. 984–991.
[3] Michael Hagenau, Michael Liebmann, and Dirk Neumann. 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems* 55, 3 (2013), 685–697.
[4] Zhongyi Hu, Yukun Bao, Raymond Chiong, and Tao Xiong. 2017. Profit guided or statistical error guided? A study of stock index forecasting using support vector regression. *Journal of Systems Science and Complexity* 30, 6 (2017), 1425–1442.
[5] Mathias Kraus and Stefan Feuerriegel. 2017. Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems* 104 (2017), 38–48.
[6] Christopher D Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* MIT press.
[7] Vladimir Vapnik. 2013. *The nature of statistical learning theory.* Springer.