Stock Market Prediction using Financial News Articles on Ho Chi Minh Stock Exchange

Duc Duong
University of IT - VNU HCMC
Ho Chi Minh, Viet Nam
ducdm@uit.edu.vn

Toan Nguyen
University of IT - VNU HCMC
Ho Chi Minh, Viet Nam
kentuit@gmail.com

Minh Dang
University of IT - VNU HCMC
Ho Chi Minh, Viet Nam
danglienminh93@gmail.com

ABSTRACT

In this paper, we examined the effects of financial news on Ho Chi Minh Stock Exchange (HoSE) and we tried to predict the direction of VN30 Index after the news articles were published. In order to do this study, we got news articles from three big financial websites and we represented them as feature vectors. Recently, researchers have used machine learning technique to integrate with financial news in their prediction model. Actually, news articles are important factor that influences investors in a quick way so it is worth considering the news impact on predicting the stock market trends. Previous works focused only on market news or on the analysis of the stock quotes in the past to predict the stock market behavior in the future. We aim to build a stock trend prediction model using both stock news and stock prices of VN30 index that will be applied in Vietnam stock market while there has been a little focus on using news articles to predict the stock direction. Experiment results show that our proposed method achieved high accuracy in VN30 index trend prediction.

Categories and Subject Descriptors

H.3.1 [**Information Storage And Retrieval**]: Content Analysis and Indexing—*dictionaries, linguistic processing*

Keywords

Stock market, Text mining, SVM, Prediction

1. INTRODUCTION

Stock market prediction is always a challenging task because it is highly volatile and dynamic. Many methods have been proposed to forecast the future direction of the stock market. One of the most significant factors impact on human's actions in stock market is news articles. Recently, the number of online news rocketed making investors difficult in reading and updating all the latest information that may affect their stocks. So automated systems should be developed and hopefully it will be useful for investors. Take stock trends as an example, if the direction of selected stock were predicted to be "up" in the next 24 hours, investors shares by using that information helping them make a wise trading action.

For years, the stock market prediction has just depended on historical market data. Researchers applied lots of algorithms such as: Moving average [8], Multiple Kernel Learning [14], Support Vector Machines [7] and other techniques to analyze the stock market behavior. Although they had a promising result, these approaches are difficult to predict accurately because researchers tried to predict the stock market from historical prices with such a random behavior of the stock market while there are no justification for it. Some real life events may cause good or a bad effects on the stock market. For example, if the price of gasoline dropped sharply it would affect investors to sell all shares they hold in petroleum securities. As a result, the stock prices in petroleum securities will go down as a result to reflect bad events.

In Vietnam, the Ho Chi Minh Stock Exchange (HoSE) has just been found since 2000, stock trend prediction by financial news has not been researched yet. Moreover, Vietnamese language has a different structures from English language [2]. That is our motivation to do this paper and propose a stock trend prediction system for Vietnam stock market with financial news and daily stock prices as an input.

This article is arranged as follows: section 2 discusses about previous works. Section 3 describes our proposed system. Section 4 explains the way we collected data. Section 5 shows the experiment results and analysis. Section 6 delivers our conclusions with a brief discussion about what we need to do next.

2. RELATED WORK

Before digging any deeper to our approach, it is worth answering whether the stock market prediction is feasible or not? According to the Efficient Market Hypothesis (EMH): "In the financial market, profit opportunities are exploited as soon as they arise, hence stock prices, historical data and general information as well as private information at any moment which will make it extremely difficult to predict". However, as pointed out in [7, 8], it is possible to forecast the stock market. In fact, it takes time for the market to adjust itself to the incoming news. It will be more profitable to generate an action signal (buy, sell) in corresponding to the market news than to accurately predict the future prices

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IMCOM '16, January 04-06, 2016, Danang, Viet Nam

© 2016 ACM. ISBN 978-1-4503-4142-4/16/01…$15.00

DOI: http://dx.doi.org/10.1145/2857546.2857619

of the stock.

Different dimensions have to be considered when we classify the stock market:

- Input: The first approach is based on historical market prices and use technical analysis to predict the stock market, the second approach is based on using news content, however the combination of both methods will be our choice to increase the accuracy of the system.
- Goal: Aim of prediction varies from predicting the future stock prices to minimize the volatility of the prices or the market trends. Market trends are the general direction of the stock prices: upward or downward. Market volatility is the indicator of the fluctuation. A higher volatility means the higher fluctuation of the corresponding stock prices
- Time span: A horizon of time needs to be considered. It can be a short-term or a long-term prediction. Short-term prediction lasted from 5 minutes to 1 day after news was published while long-term prediction started from week, month and lasted longer. This paper sticks with the short-term prediction because when the important news about the company are published, short-term investors will monitor these news in a short time, then they will decide to buy, sell or keep shares they hold based on how they think it will affect the stock.

So far, there have been two different approaches in labeling documents. The first approach is to assign a class to the article manually by expert's opinions about the content of the article. Although the successful rate is a bit higher by using this method, so many articles will be relatively hard by using only human effort. On the contrary, the second approach attaches a label to articles automatically according to their effect on stock prices. This method is less accurate than the first method because there were many different reasons stock price's change does not indicate the actual label of the article. For example, although the article is positive, global finance crises may cause a drop in the stock price. Besides, we applied our system in Vietnam stock market where news are not accurately reflecting all the information about the company operations as in other countries. News in Vietnam can be manipulated by many individuals using online sources to spread rumors that make the stock price decreased or increased according to their wishes so it is vital to find reliable sources of information.

Recently, a new approach has been proposed to increase the accuracy of predicting stock trends using news articles. [12, 4] applied sentiment dictionary-a dictionary used for estimating the effect of each word whether positive or negative to news articles in their prediction model and they proved that the accuracy has been improved.

Dimension reduction techniques can be classified into Feature Extraction (FE) and Feature Selection (FS) [10]. FS algorithm selects a subset of the most representative features from the original feature space. FE algorithm transforms the original features spaces to smaller features spaces to reduce the dimension. Though the FE algorithm is proved to be effective for dimension reduction, FE algorithm is not optimal to the high dimension of data set in the text domain due to their high computation. Thus FS algorithm is more popular for real life text data dimension reduction problems.

3. PROPOSED METHOD

3.1 Documents preprocessing

All the news articles are collected in HTML extension which contained many unnecessary tags of HTML format so we have to eliminate them first then we save exported content in plain text format for further processing.

Now each document contains lots of sentences, our next step is tokenizing each sentence into words. We use the state-of-the-art Vietnamese word segmentation tool vnTokenizer [6]. This tool is proved to achieve more than 90 percent accurate in tokenizing Vietnamese sentences. The collection of words will be the input for the next step

In the final step of this phase, the relevant words are extracted from the document. As usual, all words as well as numbers are considered to be features. After the set of features is extracted, stop words in Vietnamese language are removed; it will at least improve the efficiency of the target system and reduce the system resources. For this purpose, a list of Vietnamese stop words has been gathered and used as a stop list dictionary (about 900 words collected manually).

3.2 Documents Labeling

Based on the approaches so far: 2 or 3 classes are defined for predicting market directions. The goal of processing news generally is to classify the news into two classes either good news or bad news regarding the selected stock [7]. Sometimes, this classification is extended and another category indicating neutral news is added [5]. In our paper, we chose 3 classes approach: rise, drop and neutral to reflect all the possible direction of the stock market.

3.3 Building sentiment dictionary

After learning the effect of this dictionary in improving the accuracy. We have built our own dictionary for financial news articles. Firstly, we downloaded Vietnamese dictionary in plain text format containing over 75,000 words. Secondly, we used vnTagger tool [9] for tagging words (noun, verb, adjective,...) in that dictionary, then we filtered out only adjective and verb (increase, strong,...). Lastly, our system iterated through all the news that has been labeled, we count the words in the dictionary occur in the document with positive and negative class and we applied it in below formulas to calculate positive and negative point of each word in the dictionary.

$$t_{\mathrm{p,wi}} = \frac{|P|}{|P| + |N|} \tag{1}$$

$$t_{\mathrm{n,wi}} = \frac{|N|}{|P| + |N|} \tag{2}$$

With $t_{\mathrm{p,wi}}$ is the point representing the positive impact of word wi, $t_{\mathrm{n,wi}}$ is the point representing the negative impact of word wi. $|P|$ is the number of documents wi appear labeled positive, $|N|$ is the number of documents wi appear labeled negative.

We then browsed through all the articles; words in the dictionary that do not appear in any articles will be eliminated to reduce the processing time.

3.4 Features weighting

Each document is represented as a multidimensional vector based on all selected features as described in document

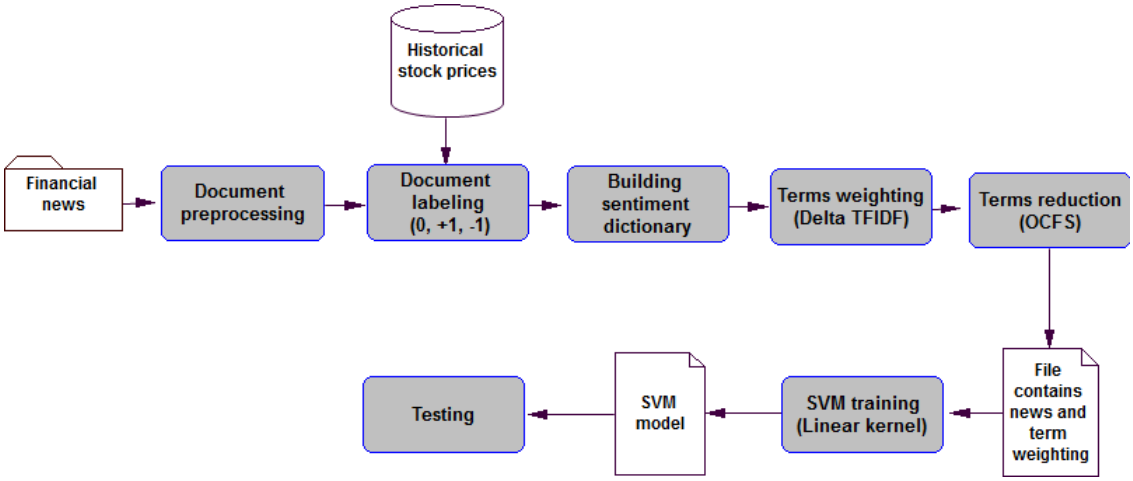

Figure 1: An overview of our stock trend prediction process

pre-processing step. Several methods of feature weighting have been proposed by researchers in text mining field. We used delta TFIDF method [11] instead of normal TFIDF, the goal was to increase the importance of the word unevenly distributed between positive and negative class, reduce the importance of the word evenly distributed between positive and negative class. Below is the formula for the algorithm:

$$V_{t,d} = C_{t,d} * \log_2(\frac{|P|}{P_t}) - C_{t,d} * \log_2(\frac{|N|}{N_t})$$

$C_{t,d}$ is the number of occurrences word t appear in the document d, P_t is the number of documents with positive class word t appears, $|P|$ is the number of documents in positive class, N_t is the number of documents with negative class word t appears, $|N|$ is the number of documents in negative class, $V_{t,d}$ is the weight of t in document d.

3.5 Terms reduction

Term reduction main aim is to decrease the number of features that need to be processed by the system. There have been many researches on finding the best feature selection method in text classification [13] such as: MI (Mutual Information), IG (Information Gain), GSS (GSS coefficient), CHI (Chi-square), OR (Odds Ratio), DIA association factor or RS (Relevancy score).

Recently, the work in [15] had shown that OCFS gives the state-of-the-art performance of FS algorithm in dimension reduction. The main idea of OCFS is:

- Calculate centroid m, i=1,2,...,c for each category of training corpus.
- Calculate centroid m for all training categories.
- Calculate the score for each term i-th.
- Choose K terms which have the highest score.

We chose OCFS as our dimensional reduction method.

3.6 SVM

We used support vector machine [1] as our machine learning method to classify financial news articles. SVM is proved to be one of the most efficient techniques for data classification. SVM is based on decision boundary, which separates sample of different classes. A good decision boundary which separates samples of different classes, it must be far away from the samples of all classes, which are separated. In this paper, we used linear kernel in SVM classification because the number of features were large so we do not need to map data to a higher dimensional space. Moreover, nonlinear mapping did not improve the performance. Linear kernel was good enough; we only needed to search for the suitable parameter C. Although SVM is considered easier to use, users who do not familiar with it often get unsatisfactory results in their first implementation. LibLinear [3] is a library for linear kernel support vector machines (SVM). Its goal is to help users to easily use linear SVM and customize it to serve their purpose.

4. DATA PREPARATION

We gathered stock news automatically from popular financial websites such as: vietstock.vn, hsx.vn, hsn.vn between May 1st, 2014 and April 30th, 2015 by using our web crawler tool. As the result, we collected 1884 different news article files. In our work, we selected only news related to companies in VN30 Index (BVH, CII, CSM, DPM, DRC, FLC, FPT, GMD, HAG, HCM, HPG, HSG, HVG, IJC, ITA, KBC, KDC, MBB, MSN, OGC, PPC, PVD, PVT, REE, SSI, STB, VCB, VIC, VNM, VSM) because VN30 index is announced by Ho Chi Minh stock exchange and based on three criteria: market capitalization, free-float ratio and the transaction value; includes shares of 30 company listed in Hose which has the highest capitalization and liquidity. Daily stock prices in the same period were collected manually from cophieu68.com. We divided the stock news into three samples in order to make it easy to compare the result that depended on time and number of articles: the first sample contained news from January 2015 to April 2015, the second sample contained news from September 2014 to April 2015, the final sample contained news from May 2014 to April 2015 as in table 1.

5. EXPERIMENT RESULTS

Confusion matrices, precision, recall, F-measure and accuracy were used to evaluate the proposed model. In confusion

Table 1: Number of Articles by Samples

Sample	Number of articles		
	Training sets	Testing sets	Total
One (4 months)	901	386	1287
Two (8 months)	1068	457	1525
Three (12 months)	1319	565	1884

matrices, T_P, T_N indicate the right classification for the corresponding class and F_P, F_N indicate the false classification for the corresponding class.

Table 2: Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	T_P	F_N
	Negative	F_P	T_N

Accuracy is the proportion of true positive (T_P) and true negative(T_N) in all the test data. Precision is defined by the true positive (T_P) against both true positive (T_P) and false positive (F_P). Recall is defined as the proportion true positive (T_P) against both true positive (T_P) and false negative (F_N). The formula is as following:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$Precision = \frac{T_P}{T_P + F_P}$$

$$Recall = \frac{T_P}{T_P + F_N}$$

Regarding which method gained the highest result for our model, we performed different approaches and checked the performance of the model in each approach, we carried out performance comparison of the three approaches for term weighting: the first approach used TFIDF, the second approach used Delta TFIDF and the final approach used Delta TFIDF combined our proposed sentiment dictionary. We also found out the best parameters for SVM machine learning: SVM-type is Linear SVM using linear kernel with C=0.5.

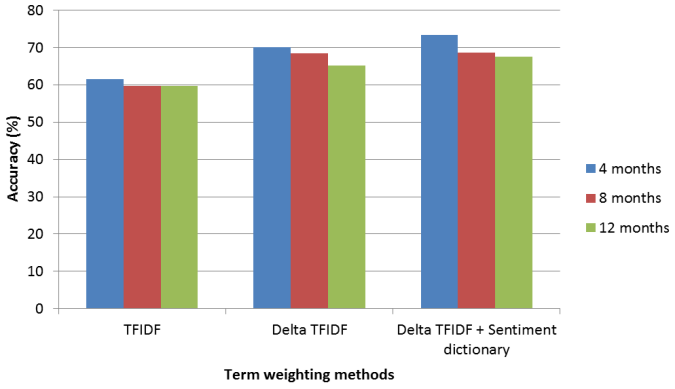

Figure 2: Comparision of term weighting techniques

As we see in the figure 2, in each time sample the accuracy of Delta TFIDF and sentiment dictionary are the highest compared to TFIDF and Delta TFIDF alone. In addition, the accuracy of time sample 1 (4 months) on each term weighting method are higher than other time samples because more and more news will lead to an unpredictable noise in prediction

Next, to confirm our theory about the noise, we performed a test on three different time samples: 4 months, 8 months, 12 months to find out the performance of the system when we increased the time span and number of articles.

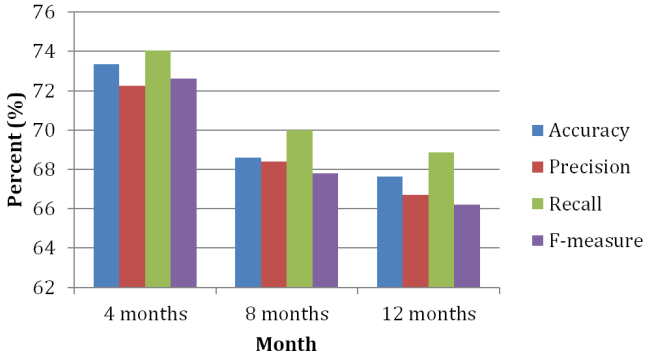

Figure 3: Term weighting techniques by samples

Figure 3 shows the accuracy of sample 1 is the best with accuracy 73% and recall 74% while in sample 2, the accuracy falls down to 68.6% and in sample 3 the accuracy is 67.6%. We have this result because the larger the number of articles the more precisely the accuracy is. In general, the accuracy is always above 60%, the fluctuation in accuracy between sample 2 and 3 is due to the noise in news articles we gathered.

To prove the possibility to predict stock trend in real scenario, we used history stock price of VN30 index in April 2015. After that, we gathered news from the same time to apply to our model. In figure 4, we saw positive trend (+1) and negative trend (-1) represented by the line below the price curve, it showed the predicted trends compared with the actual price of the stock in VN30 index. With accuracy 78.9%, the experimental result showed that the stocks trend prediction through the financial news had a high correlation with the actual price.

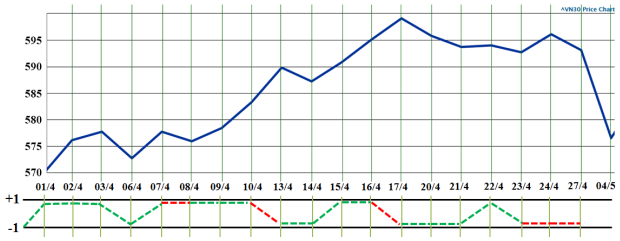

Figure 4: VN30 price chart with our model

In addition, we also tried to predict the trend of individual stock in VN30 index. We chose 5 stocks which had the highest affect among other stocks to the VN30 index: EIB, MSN, STB, VIC, VNM from March 2th, 2015 to March 13th,

Table 3: Direction of Five Stocks Ticker in VN30 Index

	EIB			MSN			STB			VIC			VNM		
	open	close	class	open	close	class	open	close	class	open	close	class	open	close	class
02/03	13.2	13.1	1	85.5	86.5	1	19.5	19.4	1	49.6	49.9	1	108	107	-1
03/03	13.1	13.1	1	87	90	1	19.5	19.4	-1	49.9	52	1	107	108	1
04/03	13.1	13.2	1	91	89.5	1	19.4	19.5	-1	52	51.5	1	108	109	-1
05/03	13.1	13.1	-1	89	88.5	1	19.5	19.3	1	51	51	1	109	108	1
06/03	13.1	13.2	1	88	88	1	19.3	19.6	1	50	49.9	-1	108	107	-1
09/03	13.2	13.2	-1	88	88	-1	19.6	20	-1	49.9	49.7	-1	107	107	1
10/03	13.3	13.3	1	86	87.5	1	20.1	19.8	1	49.5	49.7	1	107	108	1
11/03	13.2	13.2	-1	88	87	1	20.3	20.4	1	49.7	49.3	-1	108	108	1
12/03	13..1	13.2	1	89	87.5	1	20.4	20.4	1	49.3	49.6	1	108	109	1
13/03	13.2	13.2	1	88	87	-1	20.4	20	-1	49.6	49.6	1	108	108	-1

2015. We gathered 50 news for each stock by day. The result in table 3 showed that we achieved 76% accuracy in predicting the individual stock which was a promising result.

After that, we created a chart for each stock in table 3 using predicted direction.

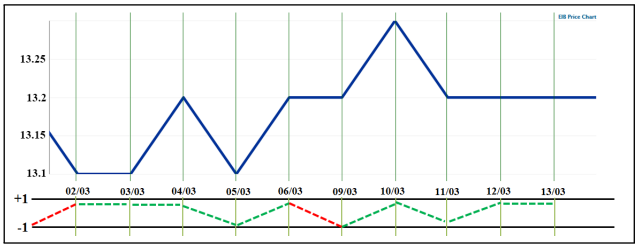

Figure 5: Trend prediction and EIB price chart

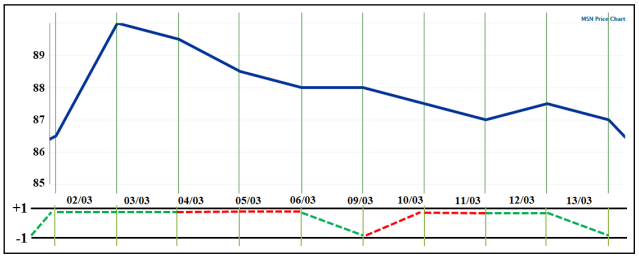

Figure 6: Trend prediction and MSN price chart

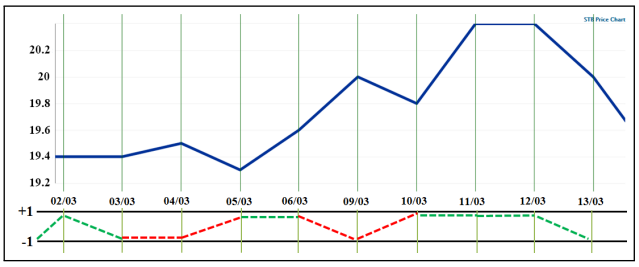

Figure 7: Trend prediction and STB price chart

As we see from figure 5 to figure 9. EIB ticker in figure 5 and VNM ticker in figure 9 achieve 80% accuracy. Especially VIC ticker in figure 8 achieve 90% accuracy. Both MSN, STB tickers in figure 6 and figure 7 achieve 60% accuracy,

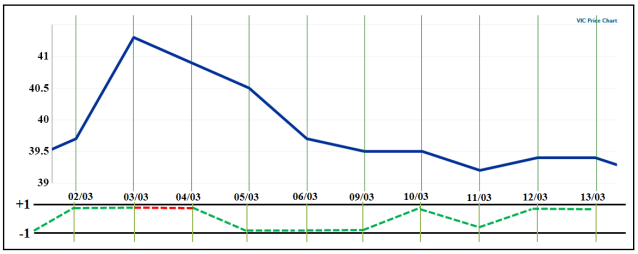

Figure 8: Trend prediction and VIC price chart

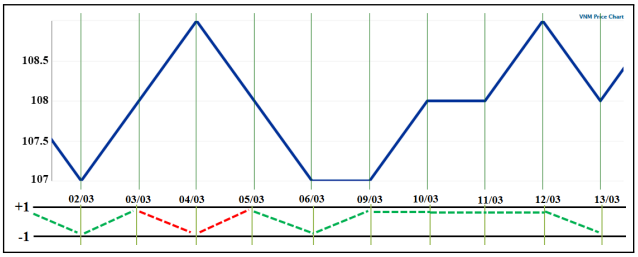

Figure 9: Trend prediction and VNM price chart

these two tickers are lower than others but it is still higher than random prediction. By predicting 5 tickers in VN30 index, we see the promising result in predicting the right direction of stock price.

Through all experimental sections above, our proposed model is proved to be possible to predict stock trends based on financial news and stock prices. In addition, by combining several methods such as: delta TFIDF, sentiment dictionary. It is clear that the accuracy of the system is greatly improved.

6. CONCLUSION

In our paper, we have proved the correlation between financial news and stock prices in VN30 index. We have achieved quite a high accuracy in both VN30 trend prediction and moreover five stocks in VN30 index. However, the success ratio of our system would increase if we could find a reliable source of news that reflects the actual stock market trend in Vietnam. In the future, we will improve the reliability of the program based on comparing SVM algorithm with other statistical methods such as Nave Bayes.

7. REFERENCES

[1] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[2] D. Dien, H. Kiem. Vietnamese word segmentation. In *NLPRS*, volume 1, pages 749–756, 2001.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[4] Y. Gao, L. Zhou, Y. Zhang, C. Xing, Y. Sun, and X. Zhu. Sentiment classification for stock news. In *Pervasive Computing and Applications (ICPCA), 2010 5th Int Conference on*, pages 99–104. IEEE, 2010.

[5] G. Gidofalvi and C. Elkan. Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*, 2001.

[6] N. T. M. Huyên, A. Roussanaly, H. T. Vinh, et al. A hybrid approach to word segmentation of vietnamese texts. In *Language and Automata Theory and Applications*, pages 240–249. Springer, 2008.

[7] M. Y. Kaya and M. E. Karsligil. Stock price prediction using financial news articles. In *Information and Financial Engineering (ICIFE), 2010 2nd IEEE Int Conference on*, pages 478–482. IEEE, 2010.

[8] S. Lauren and S. D. Harlili. Stock trend prediction using simple moving average supported by news classification. In *Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of*, pages 135–139. IEEE, 2014.

[9] P. Le-Hong, A. Roussanaly, T. M. H. Nguyen, and M. Rossignol. An empirical study of maximum entropy approach for part-of-speech tagging of vietnamese texts. In *Traitement Automatique des Langues Naturelles-TALN 2010*, page 12, 2010.

[10] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics, 1992.

[11] J. Martineau and T. Finin. Delta tfidf: An improved feature space for sentiment analysis. In *ICWSM*, 2009.

[12] P. Meesad and J. Li. Stock trend prediction relying on text mining and sentiment analysis with tweets. In *Int and Communication Technologies (WICT), 2014 Fourth World Congress on*, pages 257–262. IEEE, 2014.

[13] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[14] A. K. Sirohi, P. K. Mahato. Multiple kernel learning for stock price direction prediction. In *Advances in Engineering and Technology Research (ICAETR), 2014 Int Conference*, pages 1–4. IEEE, 2014.

[15] J. Yan, N. Liu, B. Zhang, S. Yan, Z. Chen, Q. Cheng, W. Fan, and W.-Y. Ma. Ocfs: optimal orthogonal centroid feature selection for text categorization. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. ACM, 2005.