Improvement Methods for Stock Market Prediction using Financial News Articles

Minh Dang, Duc Duong
University of Information Technology – VNU HCMC
Email: {minhdl, ducdm}@uit.edu.vn

***Abstract*—News articles serve the purpose of spreading company's information to the investors either consciously or unconsciously in their trading strategies on the stock market. Because of the immense growth of the internet in the last decade, the amount of financial articles have experienced a significant growth. It is important to analyze the information as fast as possible so they can support the investors in making the smart trading decisions before the market has had time to adjust itself to the effect of the information. This paper proposes an approach of using time series analysis and improved text mining techniques to predict daily stock market directions. Experiment results show that our system achieved high accuracy (up to 73%) in predicting the stock trends.**

***Index Terms*—Stock market, Text mining, SVM, Prediction**

I. 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 directions of the stock market. Moreover, news is one of the most significant factors impacting people's reaction in the stock market. Recently, the number of online news have rocketed which make it hard for the investors to cover all the latest information. As a result, many kind of automated systems have been implemented to support the investors. Take stock trends as an example, if the the direction of the selected stock was predicted to be "up" in the next 24 hours, investors would definitely hold that stock shares to earn more profit.

For years, the stock market prediction has just depended on the historical market data. Researchers applied a variety of algorithms such as: Moving average [9], Multiple Kernel Learning [15], Support Vector Machines [8] and other techniques to analyze the stock market's behavior. Although, these researches had a promising result, they could not predict the stock market accurately because researchers tried to predict the future prices from the historical prices with such a random behavior of the stock market and there are no justification for it. Sudden events will cause instant effects on the stock market. For example, if the price of gasoline dropped sharply, it would motivate investors to sell their shares in the petroleum securities. As a result, the stock prices in petroleum securities will decrease remarkably to reflect the bad event.

In Vietnam stock market, [3] directly predicted the stock directions with a high accuracy. However, they do not applied their findings in a real trading environment to convince the investors to apply their methods.

This article is arranged as follows: section 2 discusses about previous works the section 3 describes how our system works. After that, section 4 explains the way we collected data while section 5 shows the evaluation and result analysis. Finally, section 6 delivers our conclusions with a brief discussion about what we are going to do next.

II. Related Works

Before taking a closer at our approach, it worth answering whether the stock market prediction is feasible or not?. According to the Efficient Market Hypothesis (EMH)[5], [3]: "In the financial market, opportunities are exploited as soon as they arise. On the one hand, plenty sources of the information such as the stock prices, historical data and company's information make it extremely hard to predict accurately", but on the other hand as pointed out in [8], [9], 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 rather than accurately predict the future prices of the stock. Different dimensions have to be considered when we predict the trends of the stock market:

- Input: The first approach is based on the historical prices and technical analysis to predict the stock market, the second approach is relied on the news articles. The combination of both methods will be our choice to increase the accuracy of the system.
- Goal: The prediction's goal varies from predicting the future stock prices to minimize its volatility. The market trends are the general directions of the stock prices: upward, downward or unchanged. The market volatility is an indicator of the fluctuation of stock prices. A high volatility means a high fluctuation of the corresponding stock prices.
- Time span: A time horizon also needs to be considered. It can be for a short-term or a long-term prediction. The short-term prediction lasted from 5 minutes to 1 day after the news were published while the long-term prediction started from weeks, months and even longer. In this paper, we will apply the short-term prediction because whenever the important news are published, the investors will update these news in a short time then they will decide to buy, sell or hold their shares based on how they think they affect the stock?.

978-1-5090-2100-0/16/$31.00 ©2016 IEEE

There have been two different approaches in labeling the documents. The first approach is to assign a class for each article manually by the expert's opinions. Although, the success rate is high, the large number of articles in the dataset is relatively hard by just using human effort. The second approach is to label the articles automatically by their effect on the stock market. The later is less accurate than the former because the stock price's changes does not indicate the actual label of the article. For example, although the article is positive, global finance crises cause a drop in the stock prices. It is vital to find a reliable source of information.

One important branch of text mining is sentiment analysis, which is also referred as opinion mining. This technique explores the sentiment value of a written text. It is used to categorize text documents into a set of predefined sentiment categories (e.g. positive or negative sentiment categories) or it can be used to give the text a point on a given scale (e.g. a movie review is score on a grade from one to ten). Sentiment analysis is appropriate when applying the text mining in analyzing news articles. This is because positive news articles should have a higher probability of positively influencing the stock price, while the opposite is true for negative articles. [13], [5] applied sentiment dictionary in their news prediction model and proved that the prediction accuracy has been greatly improved.

Dimension reduction techniques are classified into Feature Extraction (FE) and Feature Selection (FS) [11]. FS algorithm selects a subset of the most representative features from the original feature space whereas 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 dataset in the text domain due to their high computation. As a result, the FS algorithm is more popular for real life text data dimension reduction problems.

III. PROPOSED METHOD

Initially we scrawl the articles from the online website then we extract the file's content in plain text format. Later we need to pre-processing all the news in the collection in order to have an optimized dataset. Next we label each news into a specific class of positive, negative or neutral by using the stock prices. Then the dataset will go through natural language processing phase including: terms weighting, terms selection. The final step is the training and testing by SVM. Below is the detailed description of each step in our model.

A. Documents preprocessing

All news articles are gathered in HTML extension which contained many unnecessary tags so we have to eliminate them first. After that, we put the exported content in plain text format. The final step in this phase is pre-processing the document as follows:

- Tokenize: The textual representation schemes based on "Bag-of-words" approach is considered as standard in prior researches because of its promising result. Basically, the text in the document is divided into a sequence of tokens as one single word which becomes one dimension of the feature vector.
- Stopword removal: All the stop words as well as numbers that do not have any useful value such as "and", "or", "by",... are removed.
- Finally all the punctuations and numbers are removed.

B. Documents Labeling

Generally, the news articles are believed to cause the movements based on delta closing price within the day the articles are established. Delta closing price for a specific day is formulated as a change in closing price from the previous day P_{i1} to closing price in the day P_i:

$$\Delta P = P_{\mathrm{i}} - P_{\mathrm{i\text{-}1}} \tag{1}$$

The goal of processing news is to classify news into different classes [8]. Based on the approaches so far, three classes have been defined for predicting the market directions: "Upward", "Neutral" and "Downward". If the delta price of a day_i is greater than zero, all the news articles published in the day *i* are labeled as "Upward". If the delta price is less than zero, all the news articles published in the day_i are labeled as "Downward" and if the delta price is equal to zero, all the news articles published in the day *i* are labeled as "Neutral".

$$The\ label\ of\ the\ articles = \begin{cases} Upward & \Delta P > 0 \\ Neutral & \Delta P = 0 \\ Downward & \Delta P < 0 \end{cases}$$

C. Sentiment dictionary

After learning the effect of this dictionary in improving the accuracy. We have built our own dictionary for the financial news articles. Firstly, we downloaded a Vietnamese dictionary in plain text format containing over 75,000 words. Secondly, we used the vnTagger tool [10] for tagging words (noun, verb, adjective,...) in that dictionary, then we only filtered out adjective and verb (increase, strong,...). Lastly, our system iterated through all the news that have been labeled then counted the words in the dictionary occurring in the document with positive and negative class and we applied them in below formulas to calculate the positive and negative point of each word in the dictionary.

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

$$t_{\mathrm{n,wi}} = \frac{|N|}{|P| + |N|} \tag{3}$$

With $t_{p,wi}$ is the point representing the positive impact of the word wi, $t_{n,wi}$ is the point representing the negative impact of the word wi. $|P|$ is the number of documents wi appears with the positive label, $|N|$ is the number of documents wi appears with the negative label.

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

D. Terms weighting

Each document is represented as a multidimensional vector of the selected terms as described in the pre-processing step. Several terms weighting methods have been proposed by researchers in the text mining field and we will used delta TF-IDF method [12] instead of the normal TF-IDF, the goal of this new algorithm was to increase the importance of the word which unevenly distributed between positive and negative class, reduce the importance of the word which evenly distributed between positive and negative class. Below is the formula for the algorithm:

$$V_{t,d} = C_{t,d} * \log_2(\frac{|P|}{P_t}) - C_{t,d} * \log_2(\frac{|N|}{N_t}) \quad (4)$$

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

E. Terms reduction

This paper will prefer the terms reduction algorithm for optimizing the system performance. Several researches have been done on the terms reduction methods in the text classification [14] such as: MI (Mutual Information), IG (Information Gain), GSS (GSS coefficient), CHI(Chi-square), OR (Odds Ratio) and RS (Relevancy score).

Recently, the work in [16] had shown that OCFS had the state-of-the-art performance of the FS algorithm in the dimension reduction. The main idea behind OCFS is

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

F. SVM

We used support vector machine [1] as our machine learning method to classify the 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 the samples of different classes, it must be far away from the samples of all classes, which are separated. In this paper we will use the linear kernel in SVM classification because the number of terms are large, we do not map data to a higher dimensional space because the nonlinear mapping does not improve the performance. Linear kernel was good enough and we only need to search for the 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 [4] is a library for linear kernel support vector machines (SVM). Its goal is to help the users to easily use linear SVM and customize it to serve their purposes.

TABLE I: THE NUMBER OF ARTICLES BY SAMPLES

Sample	Number of articles		
	Training data	Testing data	Total
One (a quarter)	901	386	1287
Two (two quarters)	1068	457	1525
Three (three quarters)	1319	565	1884

IV. DATA PREPARATION

We gathered the 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 a web crawler tool. As the result, we collected 1884 different articles. In our work, we selected only the news relating to companies in the 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 the VN30 index is announced by the Ho Chi Minh stock exchange and based on three criteria: market capitalization, free-float ratio and the transaction value; includes the shares of 30 companies listed in the HoSE which have the highest capitalization and liquidity. The 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 easier to compare the result: the first sample contained the news from January 2015 to April 2015, the second sample contained the news from September 2014 to April 2015, the final sample contained the news from May 2014 to April 2015 as shown in Table I.

V. EVALUATION

Confusion matrices, precision, recall, F-measure and accuracy were used to evaluate the proposed model. In the confusion 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 II: 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 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 of true positive (T_P) against both true positive (T_P) and false negative (F_N). The formula is as following:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (5)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (6)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (7)$$

We implemented our test on three different time samples: a quarter, 2 quarters and 3 quarters to examined the performance of the system when we increased the time span and the number of articles. We also figured out the best parameters for SVM machine learning: SVM-type was Linear SVM using linear kernel with C=0.5.

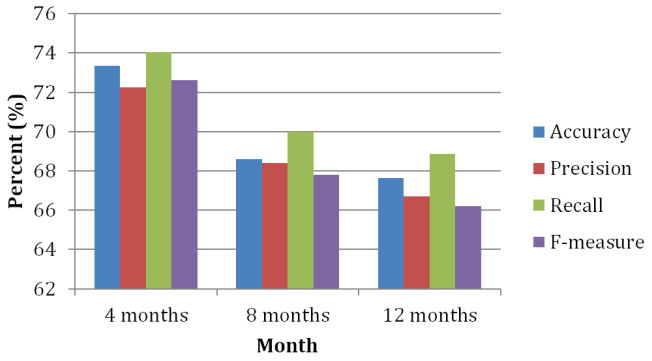

Fig. 1: The Term Weighting Techniques by Samples

Figure 1 shows that sample 1 has the highest accuracy at 73% and recall 74% while in the sample 2, the accuracy drops to 68.6% and in the sample 3 the accuracy is 67.6%. We had this result because the larger the number of the articles the better the accuracy. In general, the accuracy is always above 60%, the fluctuation in the accuracy between the sample 2 and 3 are likely due to the noise in the news articles we gathered.

To prove the possibility to predict the stock trend in the real scenario, we used the VN30 index historical stock prices in april 2015. After that, we gathered the news from the same time to apply in our model. In figure 2, we saw the positive trend (+1) and the negative trend (-1) represented by the line below the price curve, they showed the predicted trends comparing with the actual price of the stock in the VN30 index. With the accuracy of 78.9%, the experimental result showed that there was a high correlation between the stocks trend prediction and the actual prices.

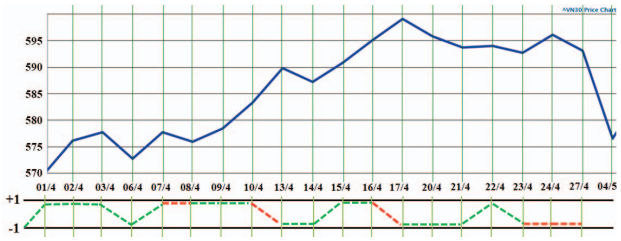

Fig. 2: Comparison of the VN30 Price Chart and Our Model Prediction

In addition, we also tried to predict the trend of individual stock in VN30 index. We chose 5 stocks which had the best technical indicators such as: EPS, PE, ROA, ROE among other stocks in the VN30 index: EIB, MSN, STB, VIC, VNM from March 2th, 2015 to March 13th, 2015. We gathered 50 daily news for each stock. The result in Table III showed that we got 64% accuracy in predicting the individual stock which was quite a promising result.

As we see from Table III. EIB ticker and VNM ticker achieve 80% accuracy. Especially VIC ticker achieve 90% accuracy. Both MSN, STB tickers achieve 60% accuracy, these two tickers are lower than others but they are still higher than a random prediction. By predicting 5 tickers in the VN30 index, we see a promising result in predicting the right direction of the stock prices.

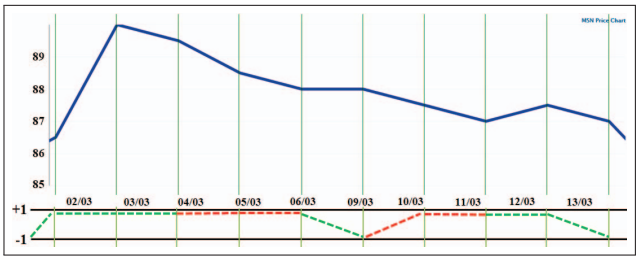

Fig. 3: Trend Prediction and MSN Price Chart

Next, a stock market simulation trading system is conducted to evaluate the profitability of the system under the real conditions. The initial investment is assumed to be 100 million VND and each transaction (buy/sell) is charged 0.25% of the trading money. The fortnight data period from March 02nd, 2015 and March 13th, 2015 are used in this test. Figure 3 shows the price movements of MSN during the period. Furthermore, in order to avoid excessive transaction charges that will result from the frequent operations, each day is limited to only one transaction (buy/sell). It operated under the following rules:

- If there are no news release, do nothing.
- If the news prediction is neutral for that day, do nothing.
- If the share has been bought and the news prediction is negative for that day, the share will be sold at the opening price
- If the share has been bought and the news prediction is positive, do nothing
- If the share has been sold and the news prediction is negative, do nothing.
- If the share has been sold and the news prediction is positive, buy at the opening price.

Table IV describes detailed transaction of the system. Our system have done 4 trades in total within two weeks based on the prediction model. From the initial 100 million VND, in the end of the month we earned 105,314,500 VND so the profit was over 5 million VND.

Through all the experimental sections above, our proposed model proved that it was possible to predict the stock trends using the financial news and stock prices. In addition, by combining several methods such as: delta TFIDF, sentiment dictionary and removed the weak stock tickers by applying the technical indicators, we did improve the performance and accuracy of our system.

TABLE III: DIRECTION OF THE FIVE STOCKS 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

TABLE IV: DETAILED TRADING ACTIONS

Date	Type	Shares hold	Money left	Open price
03/02	Buy	1169	50,500 VND	85,500 VND
03/09	Sell	0	102,922,500 VND	88,000 VND
03/10	Buy	1196	66,500 VND	86,000 VND
03/13	Sell	0	105,314,500 VND	88,000 VND

VI. CONCLUSION

In our paper, we have proved the correlation between the financial news and the stock prices. To achieve that, the financial news and the stock prices were gathered for careful experiment. We have achieved quite a high accuracy at 73%. Moreover, we removed the weak stock tickers in VN30 index by the technical indicators and the results proved that the performance of the system has greatly been improved. However, the success ratio of our system would be increased if we analyzed the news from more reliable sources. In the future, we will improve the performance of the system by combining the stock prices prediction and technical analysis.

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] Duc Duong, Toan Nguyen and Minh Dang. Stock Market Prediction using Financial News Articles on Ho Chi Minh Stock Exchange. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, pages 71–76. ACM, 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.