VIETNAMESE STOCK MARKET PREDICTION USING TEXT MINING

Phạm Xuân Dũng[1], Hoàng Văn Kiếm[2]

[1, 2] University of HCM City Information Technology

cnttdung@gmail.com, kiemhv@uit.edu.vn

Abstract: Stock market prediction has attracted many researches from both economists and computer scientists. Stock market prediction using text mining is the emerging field and there are some researches in this field all over the world. This is the interdisciplinary between linguistics, machine-learning and behavioral-economics.

In this paper, we propose the model which combines both numerical data and textual data and financial rules to enhance the predictability of the daily stock price trend of VN-Index. We collected textual data and numerical data from some of popular Vietnam websites for several years and use them for training and testing our model.

Keywords: text mining, support vector machines, news articles, stock market prediction.

Tóm tắt: Dự báo thị trường chứng khoán từ lâu đã thu hút nhiều nghiên cứu từ các nhà kinh tế học và các nhà khoa học máy tính. Dự báo thị trường chứng khoán sử dụng khai phá văn bản là một lĩnh vực mới nổi và đã thu hút một số nghiên cứu trên thế giới. Đây là lĩnh vực liên ngành giữa ngôn ngữ học, học máy, và tài chính hành vi.

Trong bài báo này, chúng tôi đề xuất mô hình kết hợp dữ liệu số và dữ liệu văn bản cùng với các luật về tài chính để nâng cao khả năng dự báo xu hướng của chỉ số giá chứng khoán VN-Index. Chúng tôi thu thập dữ liệu văn bản và dữ liệu số từ các website phổ biến tại Việt Nam trong một số năm và sử dụng chúng để xây dựng mô hình cũng như kiểm chứng tính hiệu quả của mô hình được đề xuất.

Từ khóa: Dự báo thị trường chứng khoán, khai phá văn bản, support vector machine, dự báo thị trường chứng khoán dựa trên tin tức.

1. Introduction

Stock market prediction is a challenging task since stock markets are unusual highly volatile, dynamic, nonlinear and chaotic. Many methods have been used for a long time to forecast the future direction of the stock market. Among these methods, Data mining and machine learning techniques, especially Support Vector Machines, Artificial Neural Networks has been used in a large number of applications to predict stock market trend based on time series data. These methods rely mainly on using structured and numerical data. However, human behaviors are always influenced which we hearing, seeing, discussing every day. One of the most significant impacts on affecting our behaviors is come from news articles. Amount of online news increasing dramatically make investors have difficult in reading and considering all of the latest information. So, automated system should be developed and will be very useful for investor. Recently, some of researches about stock price prediction based on news using text mining technique has been developed and have promising result. This is really an emerging topic in the data mining and text mining community (Fig.1).

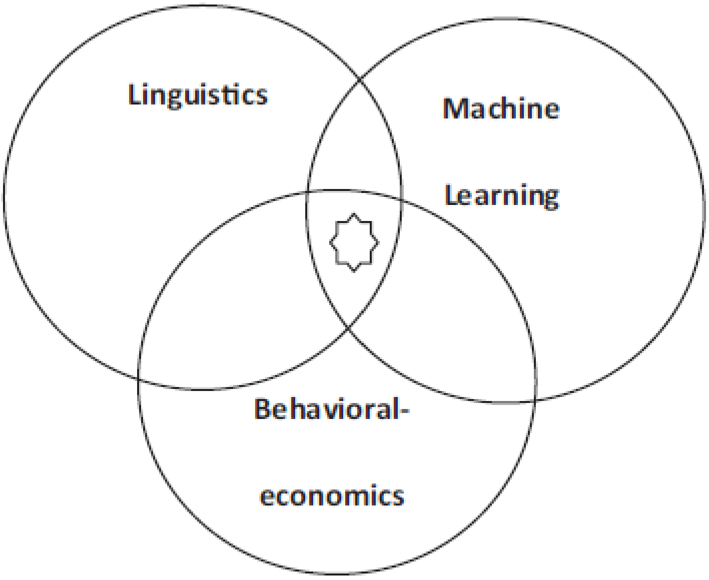

Fig.1. Interdisciplinary between linguistics, machine-learning and behavioral-economics [7]

In this paper, we propose the model use support vector machine to category textual data to predict VN-INDEX. We gather news about the stock market from some popular newspaper in Vietnam for this research.

The rest of the paper is organized as follow: Section 2 describes the related works; section 3 presents our methodology; section 4 explains the experiments and results of proposed model in this paper; section 5 is summary and conclusion.

2. Previous Works

In Viet Nam, there was some works which researched about stock price prediction such as: From economists, Lê Đạt Chí[3], using neural network for forecasting Vietnam stock market; Đặng Thị Thanh Hương[2], combine neural network and GA to predict VN- INDEX, REE, SAM in short term; Tô Nguyễn Nhật Quang[1], using GAAR(GENETIC ALGORITHM-AUTOREGRESSIVE MODEL) and ANFIS to predict VN-INDEX, REE, SAM; Phạm Thành Phước[4], using neural network to predict VN-Index; Trịnh Thanh Ngọc[5], predict orientation of market by using Twitter data, author use Support Vector Regression – SVR with data from https://twitter.com to predict stock price of Apple; Vũ Hữu Dũng[6], using data mining technique like GARCH, neural network and support vector regression to predict VNINDEX and HNXINDEX.

Stock market prediction also attract much attention over the world, in paper "*Text mining for market prediction: A systematic review"[7],* Authors review the related works that are about market prediction based on online text-mining and produce a picture of the generic components that they all have. They found that most of these systems have some of the components depicted in Fig. 2.

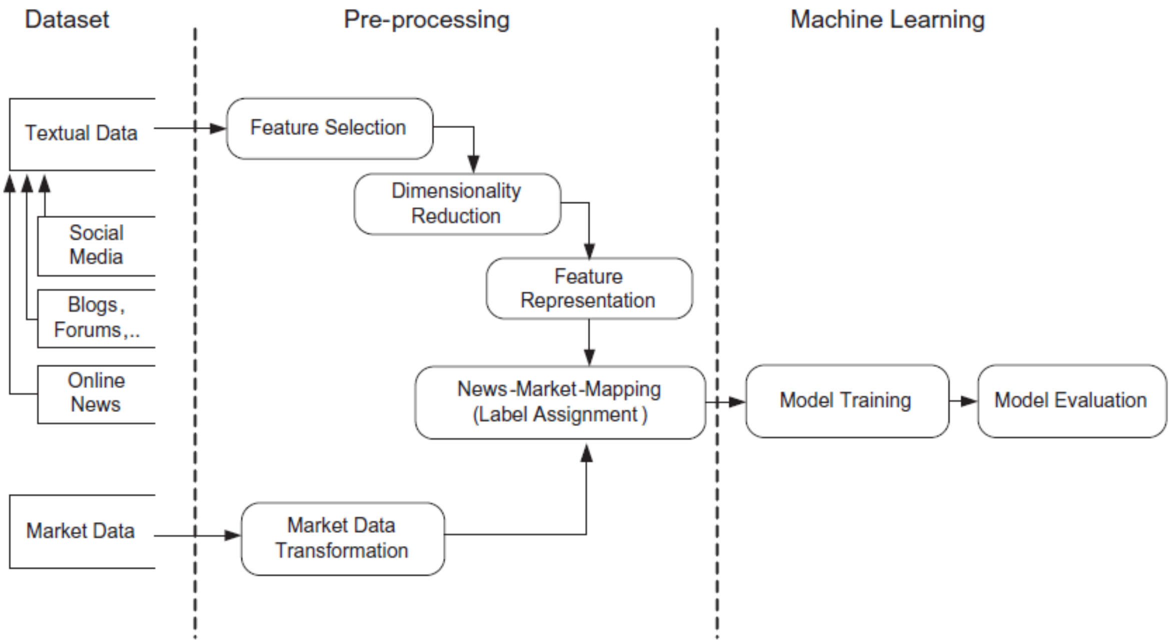

Fig. 2. Generic common system components diagram. [7]

Tien Thanh Vu, Shu Chang, Quang Thuy Ha and Nigel Collier [8], utilize Tweets data to predict daily up/down trend of stock price of Apple (AAPL), Google (GOOG), Microsoft (MSFT) and Amazon (AMZN) on The *NASDAQ* Stock Market.

Hoang T. P. Thanh, Phayung Meesad[10] combine time series data and textual data which downloaded from vietnamnews.vn to predict VN-Index for the next day. Feifei Xu [12] predicting the up and down movement of stocks by using the collective sentiments with the precision is 58.9%. Some other works such as [9], [11] also use textual data to predict stock price.

3. Approach

We propose the model as in Fig. 3.

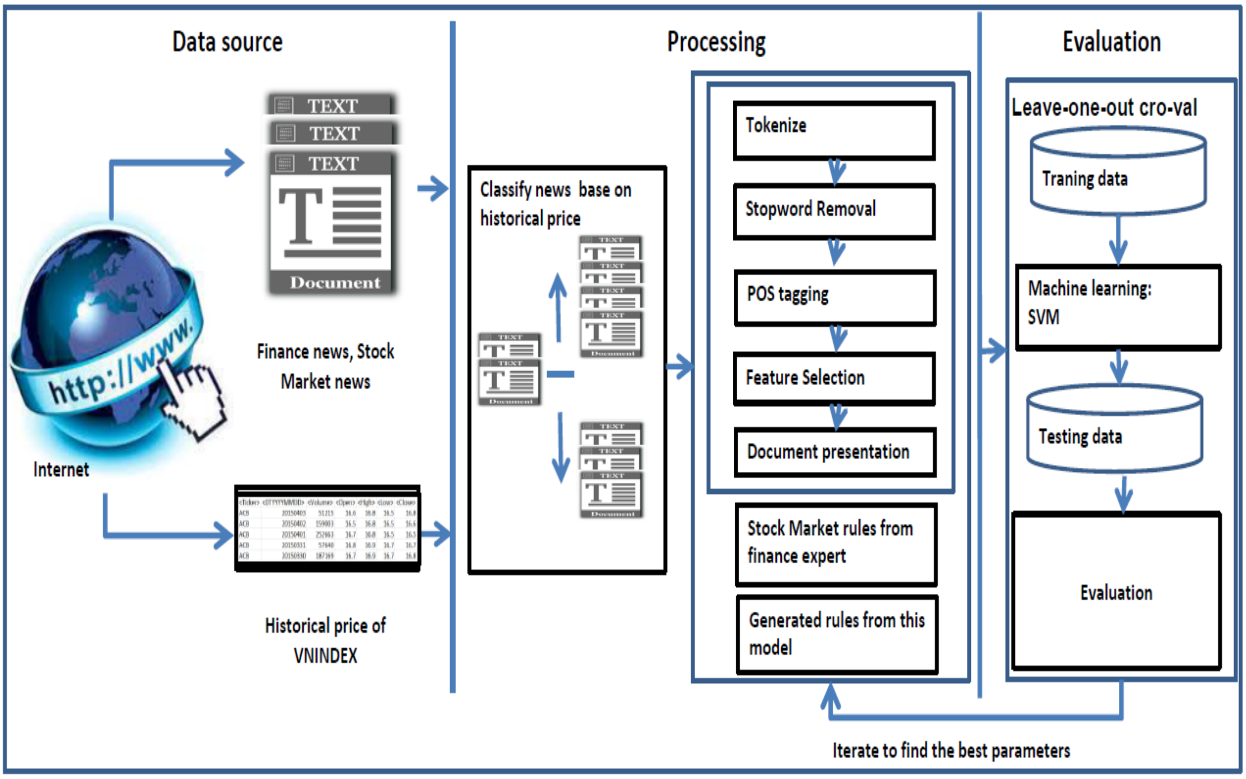

Fig. 3. Proposed model for stock market prediction using text-mining

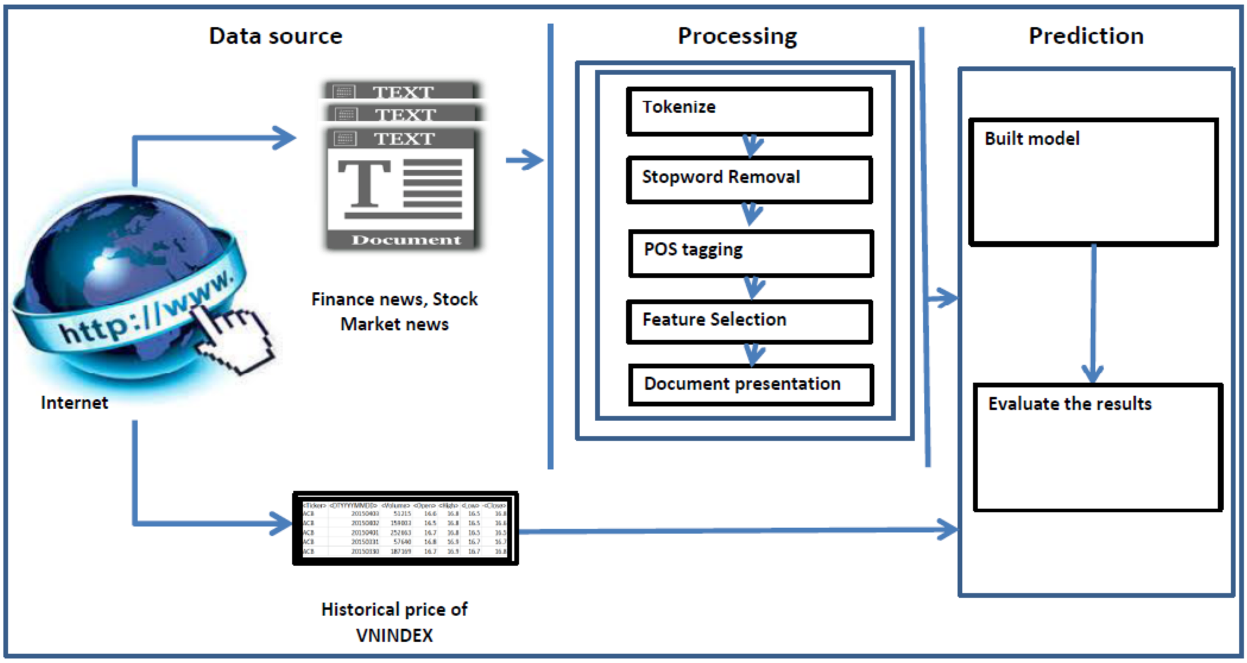

Fig. 4. Framework to evaluate the proposed model

3.1. Trend and News Alignment

After we download data from online resources, we label the news automatically to 3 categories (positive, neutral, negative) to prepare data for training phase of the proposed model.

In previous works, there were two main approaches to label article according to historical price trend. The first approach labels news articles by reading them manually. The second method is to label articles automatically according to their effects on stock price.

In this paper, we choose the second approach since we would like to extend our module to cover the increase of amount of article every day. For a specific day, the delta price ΔP is calculated as change from price of previous date P_{t-1} to price of the day P_t as formula (1).

$\Delta P = P_t - P_{t-1}$ (1)

By calculating ΔP, the class of article is determined as in (2).

Value of ΔP	Label of Articles on $Date_{t-1}$
>0	Up
=0	Neutral
<0	Down

(2)

3.2. Feature Extraction and Feature Selection

In this phase, we do following step for news:

+ Tonkenize: We utilize 2 tools **vnTokenizer 4.1.1** and **vnTagger 4.2.0** downloaded from http://www.loria.fr/~lehong/tools/vnTokenizer.php to tonenize and tag text.

+ Step word removal: We remove the stop words which do not have information such as: Á, à, ạ, á_à, a_ha, à_ơi, ạ_ơi, ai, ái, ai_ai, ái_chà, anh_ấy, anh_chàng…

+ Remove the term which occur at almost the news occurs only few times in all training data.

After the above step, we only select word which may be have the meaning of positive or negative in the news. After survey, we consider words which have following lexical tags may have high possible bring the sensitive meaning:

1. Np - Proper noun
2. Nc - Classifier
3. Nu - Unit noun
4. N - Common noun
5. V - Verb
6. A - Adjective
7. P - Pronoun
8. R - Adverb

By surveying, we also see words which belong to below lexical tags don't have much meaning to the sentiment of news.

9. L - Determiner
10. M - Numeral
11. E - Preposition
12. C - Subordinating conjunction
13. CC - Coordinating conjunction
14. I - Interjection
15. T - Auxiliary, modal words
16. Y - Abbreviation
17. Z - Bound morphemes
18. X – Unknown

3.3. Articles Representation

After step 3.2, we have list of selected features. In this step we represent each article (document) as a multidimensional vector. Each feature is regarded as a dimension of the vector.

In this paper, we use tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. This is the popular method used by researchers to present a feature as a vector.

Calculation of tf-idf of a term tj in document di: wtf-idf(tj,di) = tf (tj, di) * idf (tj) where tf(tj, di) is term frequency in a document. Idf (tj) is inverse document frequency. Inverse document frequency is calculated: idf (tj) = $\log_2$ (N / df(tj)) where N is number of documents, df(tj) is a number of documents in which term tj occurred.

3.4. Machine learning Algorithms: Support Vector Machines and Neural network

In this research, we utilize below machine learning open source framework for building the program.

- Support vector machine library for multiclass classification in machine learning framework **Accord (https://code.google.com/p/accord/)** which use SMO (sequential minimum optimization) algorithm for training the multi-class SVM.

- Feed forward artificial neural *networks* module that use *Resilient back propagation* algorithm for training the neural network in **encog-core-cs(http://www.heatonresearch.com/encog)**.

4. Experimental Results

In this section, we explain about how to get data source for this work and how to evaluate the proposed model.

4.1. Financial News Article Sources

In this paper, we develop the tool to download text data from website: http://vietstock.vn/nhan-dinh-thi-truong/nhan-dinh-ngay.htm from 20/08/2013 to 02/04/2015 with the total of articles is 313. We also collected daily stock price of VN-Index from http://www.cophieu68.vn for the purpose of align the new to historical price automatically.

4.2. Results

To evaluate the proposed model, we first use leave-one-out cross validation, we use 131 articles of the period from 23/06/2014 to 11/02/2015 to feed to the model. Because leave-one-out cross validation need N (is the number of sample in dataset) iteration, so we just extract the accuracy at some k^{th} iteration and below is the result (table 1).

Value of k(iterating value)	18	22	25	28	30
Number of correct predict	13	15	17	18	20
Percentage of correct predict	72.22%	68.18%	68.00%	64.29%	66.67%
Number of wrong predict	5	7	8	10	10

Table1: The evaluation result of proposed model (leave-one-out cross validation)

The second validation is that we use 222 articles of the period from 06/11/2013 to 17/12/2014 to training the proposed model and 25 articles of the period from 18/12/2015 to 03/02/2015 to testing the model and we get below result (table 2)

		Prediction				
		Up	Neutral	Down	Accuracy	
Actual	Up	11	4	1	68.75%	=(11/(11+4+1)) *100
	Neutral	0	0	0	N/A	
	Down	6	1	2	66.67%	=(6/(6+1+2)) *100

Table2: The evaluation result of proposed model (80% for training, 20% for testing)

5. Conclusions

In this paper, we already proposed the model for stock market prediction by using article downloaded from website specialized in stock market. The proposed model has the promising result for the prediction trend of VN-Index. This is the motivation for us to continue to enhance proposed model to get higher accuracy rate. At the other hand, the proposed model and techniques in this paper can be applied in many other applications such as in predicting foreign exchange market, sentiment analysis of product review, predicting gold price, assigning topics to news articles, e-mail filtering, customized newspapers…

We will continue enhance our model with some of below approaches and we believe that we can increase the accuracy of the proposed model:

+ Enhancing the method to label the news automatically for training data

+ Research the capability of predicting of stock price of specific companies based on news.

+ Combine news from other popular websites in Viet Nam such as http://www.thanhnien.com.vn/chung-khoan/, http://tuoitre.vn/tin/kinh-te/tai-chinh, http://kinhdoanh.vnexpress.net/tin-tuc/chung-khoan/ into the proposed model.

+ Analyze the inter-relationship of news for the group of companies.

+ Combine emerging techniques into the proposed model are active learning, learning with unlabeled data.

References

[1] Tô Nguyễn Nhật Quang(4/2007), "U*sing generic algorithm and fuzzy logic in stock market prediction.*", Master thesis, UIT-HCM.

[2] Đặng Thị Thanh Hương(12/2009), "U*sing data mining to analysis and predict movement of stock market*", Master thesis, UIT-HCM.

[3] Lê Đạt Chí(2011), "*Using neural network in economics prediction, case of Vietnamese stock market",* Doctoral thesis, University of Economics Ho Chi Minh City.

[4] Phạm Thành Phước(2013), "*Apply neural network in prediction of price of stock market at securities trading center of HCM city.*", Master thesis, Posts and Telecommunications Institute of Technology**.**

[5] Trịnh Thanh Ngọc(2013), "*Predict trends of the stock market using Twitter*", Master thesis, University of Technology, Ha Noi National University.

[6] Vũ Hữu Dũng(2013), "*Apply data mining in Vietnamese stock market movement"*
, Master thesis, University of Technology, Ha Noi National University.

[7] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, David Chek Ling Ngo(2014), "Text mining for market prediction: A systematic review", Expert Systems with Applications, Vol.41, 15 November 2014, pp.7653–7670.

[8] Tien Thanh Vu, Shu Chang, Quang Thuy Ha and Nigel Collier (2012). "An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter", IEEASMD2012, Mumbai, India, December 9, 2012, http://wing.comp.nus.edu.sg/~antho/W/W12/W12-5503.pdf

[9] Shou-Hsiung Cheng(Jul 2010), "Forecasting the change of intraday stock price by using text mining news of stock", Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, Qingdao, Vol.5, IEEE, pp. 2605 – 2609.

[10] Hoang T. P. Thanh, Phayung Meesad(2014), "*Stock Market Trend Prediction Based on Text Mining of Corporate Web and Time Series Data*", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.18 No.1, 2014.

[11] Kim-Georg Aase(2011), "*Text Mining of News Articles for Stock Price Predictions*", Master Thesis. Norwegian University of Science and Technology, Department of Computer and Information Science.

[12] Feifei Xu(Aug 2012), Data Mining in Social Media for Stock Market Prediction, Master Thesis of Electronic Commerce at Dalhousie University Halifax, Nova Scotia.

Websites:

http://vietstock.vn/nhan-dinh-thi-truong/nhan-dinh-ngay.htm

https://code.google.com/p/accord/