Stock Price Prediction Using Financial News Articles

M. İ. Yasef Kaya, M. Elif Karslıgil
Department of Computer Engineering
Yıldız Technical University
Istanbul, Turkey
miykaya@yahoo.com, elif@ce.yildiz.edu.tr

Abstract—**Stock price prediction is one of the most important issues to be investigated in academic and financial researches. Data mining techniques are frequently involved in the studies aimed to achieve this problem. In this paper we investigate predicting stock prices using financial news articles. A prediction model, finding and analyzing correlation between contents of news articles and stock prices and then making predictions for future prices, was developed. We retrieve financial news articles published in last year, and we get stock prices for same period. All articles are labeled positive or negative according to their effects on stock price. So we use price changes to label the articles. While analyzing textual data, we use word couples consisting of a noun and a verb as features instead of using single words. Afterwards, support vector machines classifier is trained with labeled train articles. Finally, classes of test articles are predicted with using the model resulted from train phase. We achieve serious success rates that prove predictive power of our system.**

Keywords-stock price prediction; data mining; text mining; text categorization; financial news

I. INTRODUCTION

There are a lot of online sources that publish financial news on the Internet to help investors for shaping their investments. Both, current and historical news about companies, economic and political events are available on these sources. Availability of this huge amount of financial data in digital media creates appropriate conditions for a data mining research.

Stock prices are determined by supply and demand of investors. The most important information that investors used to make investment decisions is financial news. But it is a hard and time consuming task to read and analyze a lot of news published on several sources. Also, impacts of news on stock prices are limited in a short time span. So, investors have not enough time to review all financial news that affect stock price.

In this study, we propose a system predicting stock price movements by analyzing financial news articles. We aim to use rich online textual information to achieve predicting stock price movements, while there are a lot of financial articles published about stocks trading on various stock exchanges. We introduce a novel approach while analyzing textual statements in news articles. In fact, we use probability statistics of occurrence of word couples composing from a noun and a verb that occur in the same sentence.

There are a lot of substantial works done on prediction of stock prices. These works are basically text categorization systems targeting to predict stock price movement by classifying financial news articles as positive or negative. Since the problem is converted to a text categorization problem, several feature selection and classification methods are used in these works. In [1], [2], term frequency – inverse document frequency technique is used as a feature selection method. Reference [3] use chi-square statistics feature selection method. Support vector machines, k nearest neighbor and naive bayes are most widely used methods for classification. In classification phase of [1], [3], [4], [5] support vector machines method is used. While [6], [7] use naive bayes, [2],[8] use k-nearest neighbor method for classification. The accuracy rates of these works are mostly below 60%. This relatively low success rates are caused by nature of stock price movements, which are result of decisions of investors, since it is hard to predict human behavior.

II. SYSTEM DESIGN

In this section, design of stock price prediction system and techniques, used in implementation of system, are explained. General architecture of the system and the associated process flow are mentioned in Section 2.1. In the following sections, the phases of system architecture are explained in detail.

A. System Architecture

Although the architecture of target system is similar to the architecture of classical text mining applications, because of financial aspect of the system, it consists of some additional phases. General architecture of the system is illustrated in Fig. 1.

978-1-4244-6928-4/10/$26.00 ©2010 IEEE

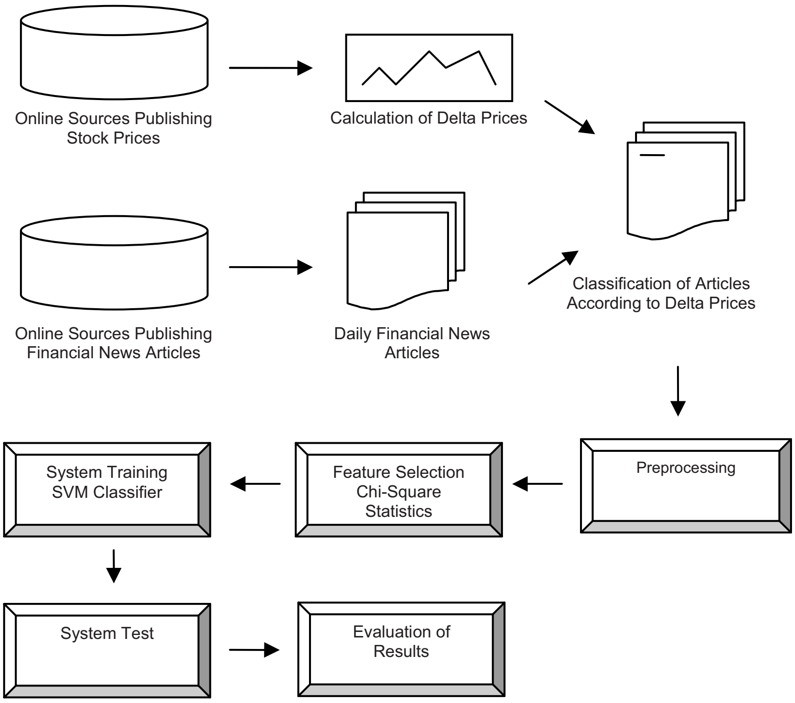
Online Sources Publishing Stock Prices
Calculation of Delta Prices
Online Sources Publishing Financial News Articles
Daily Financial News Articles
Classification of Articles According to Delta Prices
Preprocessing
Feature Selection Chi-Square Statistics
System Training SVM Classifier
System Test
Evaluation of Results

Figure 1. System architecture

In the first step, all price values and news articles, published for a specific time period, are acquired. In the next step, delta prices are calculated for each day belonging to this time period. Afterwards, the features to be used in classification of the articles are extracted. After extracting features, the ones which are most effective in classification of the articles are selected. Training of system performed on training articles according to selected features. Afterwards the system is tested with test articles using the model, which results from training phase.

B. Getting Stock Prices and Financial News Articles

Stock prices and financial news articles about stocks trading on various stock exchanges are available in several internet sites. A stock belonging to a particular company is selected for training and testing of the target system. Also a specific time period is determined for getting prices and news articles published in this same time period. This time period must have enough length to reflect success of target system. Prices of selected stock and new articles concerning it for this time period are used as dataset of the system.

C. Labeling Financial News Articles

In previous researches, generally two different approaches were used for labeling articles as positive or negative. The first approach is to read news articles and label them manually. Although success rate is higher by using this method, the number of articles used in dataset is relatively limited because it requires human effort. The second one is to label articles automatically according to their effects on stock price. While this method is used, general success rate of the overall system may be a bit low. Because the stock price changes may not indicate actual label of the article for different reasons. For example although an article is positive, global finance crisis may drop stock prices.

Since we want to constitute an extendible dataset, we prefer to label financial news articles according to their effects on stock price. Positive labeled article impacts on rising of stock price, negative one has a drop impact on stock price. We assumed that each article has an effect on stock price whether positive or negative. So, we didn't state a neutral class in the system. In order to label an article, we used delta price value for a day on which the article was published. For a specific day, the delta price ΔP is calculated as change from price of previous day P_{t-1} to price of the day P_t as illustrated in (1):

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

If delta price for a specific day is greater than zero, all articles published in this day are labeled as positive. If the delta price is less than zero, all articles published in this day are labeled as negative. We use 1 to represent positive class, -1 to represent negative class. According to this, the relationship between delta price (ΔP) and class of article is determined in (2):

$$\textit{Class of Article} \begin{cases} 1(\textit{Positive}) & \Delta P >= 0 \\ -1(\textit{Negative}) & \Delta P < 0 \end{cases} \tag{2}$$

D. Preprocessing

The most classical textual representation method for text classification systems is defining frequencies of words in texts as features. Success of this method depends on the subject of the work. For example this method is very successful for a system in defining authors of a text or classifying an e-mail as spam or not. But it is not very useful for the classification of financial news articles as having positive or negative effect on stock price. Because it is not valuable in classifying an article to know if a specific word occurs in it or not. For example to know the word "sales" occurred in an article don't give an idea whether the article has a positive or negative effect on stock price. Because, if the word "sales" takes place in a sentence like "X Company's sales increased by 20 percent in the second quarter", it has positive effect on stock price. On the other hand, if in a sentence like "X Company's profit dropped by 10 percent from last year", it would have a negative effect.

In this study, we defined word couples composed from a noun and a verb word occurring in a same sentence as features. Because it may give an idea to know that, a specific noun and a specific verb occur in a same sentence, about whether the sentence is positive or not. In previous examples, the word couple "sales" and "increased" in a same sentence gave us an idea, whether the sentence was positive or not. Occurrence of the word couple "profit" and "dropped" in same sentence says that the sentence is negative.

According to the single word feature approach, words extracted from sample sentence "X Company's sales increased by 20 percent from last year" are listed in Table 1. The words "by", "from" and "last", which are stop words, are omitted. On the other hand, word couple feature approach extracts 6 features as listed in Table 2 for the same sentence. Since there is only one verb word in the sentence, the second word of all features is same "increased". The words "X", "Company", "sales", "20", "percent" and "year" are noun words, thus they are placed in the fist word of the features. The couples including at least one stop word are removed from the feature list.

TABLE I. SINGLE WORDS AS FEATURES

	Noun word
1	X
2	Company
3	sales
4	increased
5	20
6	percent
7	year

TABLE II. WORD COUPLES AS FEATURES

	Noun word	**Verb word**
1	X	increased
2	Company	increased
3	sales	increased
4	20	increased
5	percent	increased
6	year	increased

E. Feature Selection

After defining all noun-word couples occurring in the same sentence as features, we need to select the features, which are most effective in classification. Because there are thousands of noun-word couples in the sentences of all articles. Most of them may not have any effect on classification. Also, trying to process all of them may result inadequate system performance. Thus, we must eliminate non-effective or least effective features. For example, when the Table 2 are reviewed, the couples, except "sales"-"increased", are meaningless for classification. Because they don't give any idea about whether the sentence is positive or not. So, the couple that doesn't effect classification must be eliminated. This elimination process cannot be performed manually, since there are thousands of noun-word couples. We need to use statistical techniques to determine effectiveness of features and select the most effective ones.

There are a lot of feature selection methods in the literature. Also, there are some works that compare performances of these methods. After reviewing these works, we decided to use chi-square feature selection method which is more suitable for text classification. In this method, dependence between classes and features are calculated and used to weight features. Chi-square weight indicates effectiveness of feature. Chi-square uses (3) to calculate weight of feature *i* on class *c*. *N* is the total number of documents, *P(i,c)* is number of documents in *c* containing *i*.

$$X^2(i,c) = \frac{N \cdot \left[P(i,c) \cdot P(\bar{i},\bar{c}) - P(i,\bar{c}) \cdot P(\bar{i},c)\right]^2}{P(i) \cdot P(\bar{i}) \cdot P(c) \cdot P(\bar{c})} \tag{3}$$

F. *System Training and Testing*

We selected Microsoft's stock which is trading on New York Stock Exchange. We retrieved the articles related to Microsoft's stock MSFT published in last year from www.fool.com web site. We saved 982 different news articles as text files. While we use word couples occurred in same sentences as features, the probability of feature occurrence in one particular article was very low. In order to increment possibility of feature occurrences in one sample, we composed articles that were published in the same day into one text document. So, we got 182 samples for each work day of last year.

Many classification methods can be used for document classification. We used support vector machines [9] method to classify financial news articles. SVM (Support Vector Machines) is one of the most efficient techniques for document classification. SVM is based on decision boundaries, which separate samples of different classes. A good decision boundary must be far away from the samples of all classes, which are separated.

We used SVM Light application to perform classification process. SVM Light library, an implementation of SVM, is used for classification of articles. The optimization algorithms used in SVM Light are described in [10], [11].

We used 10-fold cross validation method to perform validation of the system. In ten steps, we tested all individual samples in the dataset, iteratively.

III. Evaluation of Results

We performed validation tests and retrieved performance results of our system. We used some parameters to get best results. One of the parameters is number of the features to be used in classification. We tested the system with feature count parameter value as 100, 250, 1000, 5000 and 10000. Another parameter is kernel function of SVM. We used linear, polynomial and radial basis kernel functions, respectively.

Our validation experiments yielded accuracy of 61% best with radial basis kernel SVM and 250 features. We get 61% precision and %87 recall from our experiments. Table 3 shows the confusion matrix for SVM and our novel word couples features approach.

To evaluate our word couple features approach, we performed another test with single word features using the same dataset. We get 59% accuracy best with SVM and 5000 features. This accuracy is low compared to our novel approach. So we argued that defining word couples as features is more suitable for classification of financial articles than defining single words as features. Confusion matrix, belonging to the results of SVM and single word feature approach, is presented in Table 4.

TABLE III. Confusion matrix for word couples feature approach

		Predicted Class		
		Positive	***Negative***	
Actual Class	***Positive***	TP=90	FN=14	104
	Negative	FP=57	TN=21	78
		147	35	Total= 182

TABLE IV. Confusion matrix for single word feature approach

		Predicted Class		
		Positive	***Negative***	
Actual Class	***Positive***	TP=92	FN=12	104
	Negative	FP=63	TN=15	78
		155	27	Total= 182

Fig. 2 illustrates precision-recall curve for two different feature approaches. While blue curve represents word couple features approach, red one represents single word features approach. When precision-recall curve moves towards upper-right corner, better performance is carried since both precision and recall values are increased. According to the Fig. 2, word couple features approach has a better performance than the single word feature approach as it is positioned nearer to upper-right corner.

As we expected most of word couples, defined by our system as feature, has indicative value about article classification. In Table 5, some example word couples are given. For example while the couple consisting of "revenue" and "generates" words has a positive indicative value, the couple "loss" and "have" words has negative a indicative value.

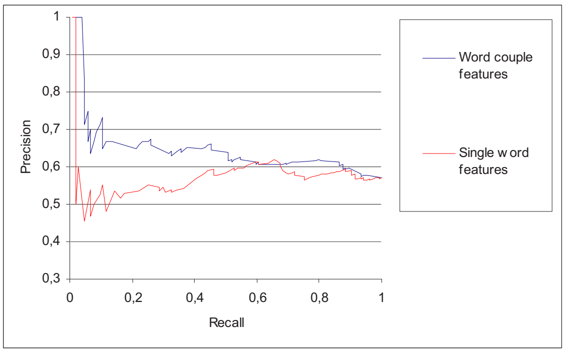

Figure 2. Precision recall curve

TABLE V. WORD COUPLE EXAMPLES

	Noun word	Verb word
1	revenue	generates
2	earnings	climbing
3	profit	growing
4	investor	offered
5	loss	have

While reviewing document based results, we noticed that some situations, classification performance were negatively affected. For example, for some days, documents include negative news, but stock prices increase in real value. Although the system classifies these documents as negative (which is true according to contents of the documents), these classifications must be accepted as false because stock price movements indicate the opposite. The same issue arises for positive news containing but negative labeled documents. Another situation is the lack of classification information for some documents. In fact, for some days, published news doesn't have any positive or negative meaning. But, stock prices change. We consider that investors make decisions in these days according to information other than the company's financial news for example global or political news. Since these kinds of news are not stated in our documents it is not possible to predict stock price movements precisely for these days.

IV. Conclusion and Future Directions

We accomplished the stock prediction system using financial news articles. Our system automatically analyzes and classifies news articles and generates recommendations for investors. We acquired 61% accuracy from our study. This accuracy rate is greater than random prediction, which has 50% accuracy. These results argue that there is a strong relationship between financial news and stock price movements.

Success ratio of our system can be increased by using more proper news articles. Because some articles in the dataset may have not been directly related to the selected stock. Online sources that we used to get articles publish some articles originally related to stocks of other companies in the same sector under our stock's category. If articles that are not directly related to the selected stock are eliminated, success of the system would increase.

References

[1] M.A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS), Big Island/Hawaii, January 2004.

[2] B. Wuthrich, V. Cho, S. Leung, D. Permunetillek, J. Zhang and W. Lam, "Daily stock market forecast from textual web data," IEEE International Conference on Systems, Man, and Cybernetics, San Diego/CA, 11-14 Oct 1998.

[3] P. Falinouss, "Stock trend prediction using news articles: a text mining apporach," Lulea University of Technology, Department of Business Administration and Social Sciences, 2007.

[4] G.P.C. Fung, J.X. Yu and H. Lu, "The predicting power of textual information on financial markets," IEEE Intelligent Informatics Bulletin, 5(1):1-10, 2005.

[5] M. Koppel and I. Shtrimberg, "Good news or bad news? tet the market decide," In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.

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

[7] P. Kroha, R. Baeza-Yates, "Classification of stock exchange news," January 2004.

[8] Y.-C. Wu, "Predicting the trend of Taiwan Weighted Stock Index with text mining techniques," Department of Information Management, National Central University, Taiwan, 2007.

[9] V.N., Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[10] T. Joachims, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 11:41-54, 1999.

[11] T. Joachims, Learning to Classify Text Using Support Vector Machines, Dissertation, Kluwer, 2002.