New Trends for Stock Market Prediction using NIPS

Jageshwer Shriwas\textsuperscript{1}, Dr. Samidha Dwivedi Sharma\textsuperscript{2}

\textsuperscript{1}M.Tech Scholer NRI Institute of Science and Technology, Bhopal(MP).
\textsuperscript{2}Professor and HOD, Department of IT, NRI Institute of Science and Technology, Bhopal (MP).

Abstract— News articles play major role in stock market prediction because these article affect the direction of decision of investors. This study investigates the prediction of stock market price changes immediately after new article publications. This process is done by automatic analysis of these new articles. In this paper we introduce a system to predict stock price trends for the time immediately after the publication of news content called NIPS (News Isolation and Prediction System). This system having four modules. First news collector module is used to collect all types of news coming from different source, second component is isolator engine in this phase all news are analyzed and categorized in three predefined classes positive, negative and neutral news. Third module is comparator in this phase we compare these categories of news to sub categories of news like finance, political and other which show us the maximum affection of news to stock price. This will help us to find more accurate result as compare to previous system.

Keywords— Classification, data mining, news article, stock price prediction.

I. INTRODUCTION

Data Mining is in essence a set of techniques that allows you to access data which is hidden in your database. It is the process of sifting through data in order to find previously unrealized patterns. This process is enabled by computer programs that are able to analyze. One area in which data mining can be useful is in financial trading markets. There are several data mining tasks, including classification, regression, clustering, dependence modeling, etc. Each of these tasks can be regarded as a kind of problem to be solved by a data mining algorithm. Therefore, the first step in designing a data mining algorithm is to define which task the algorithm will address. One of the fundamentally used algorithm is Rule based. Rule-based classifiers are classifiers in their most traditional sense. Rule-based classifiers are used to define rules that can be used to explain the results of a given training set. These classifiers begin with hypotheses of how the input attributes relate to the class attribute, and modify these hypotheses based on the data in the given training set\textsuperscript{6}.

Decision making for investors in stock market is considered to be one of the difficult tasks. There is a need for the study in data mining for shares selection with companies that has maximum growth rate. But our problem is to find out the companies which are reliable to invest in the shares with maximum profit. This is a useful approach to identify the company’s growth rate based on some of the attributes, e.g. we can examine that the “Infosys having highest growth rate with maximum volume quantity because of frequent change in high and low values in stock market”, and here we have basic property related to this example, i.e. company name, high value, low value, open, close, volume. Similarly we analyze different companies have different volumes based on their high and low values and predict the growth pattern. Thus on the basis of this scenario we can predict the reason of zero-growth, slow-growth and fast-growth items. Data mining techniques are best suited for the analysis of such type of classification, useful patterns extraction and predictions\textsuperscript{8}. Stock market analysis is one of the most interesting areas of research. Lots of investors are involved in stock market and they are all interested to know more about the future trends of market so that they gain more profit through investments. Stock information has multiple categories, i.e. stock closed price, trading volume, stock news, stock message, and expert analysis. Many studies have been done to predict the stock price by using statistics and machine learning techniques using historical stock price and trading volume\textsuperscript{5}. Classification techniques supports many algorithms. Here we used Rule based classification to categorized the news in predefined class label. Ifs-Then rules can be generated either by from decision tree or directly from the training data using or sequentially covering algorithm.

However, this paper is focus on, to improve the accuracy of stock market price prediction using NIPS (News Isolation and Prediction System). The rest of the paper arranged as follows:- Section 2 provides the review of literature. Section 3 deals with the basic architecture of System. Next section describe how data are analyzed and suitable approach for prediction. In next Section discussed the result. And last section summarize the paper.
II. REVIEW OF LITERATURE

Kim et al.[1] investigates the prediction of possible stock price changes immediately after news article publications with the help of system. This system consists mainly of four components. The first component gathers news articles and stock prices automatically from internet. The second component prepares the news articles by sending them to some document preprocessing steps and finding relevant features before they are sent to a document representation process. The third component categorizes the news articles into predefined categories, and finally the fourth component applies appropriate trading strategies depending on the category of the news article. This system requires a labeled data set to train the categorization component. This data set is labeled automatically on the basis of the price trends directly after the news article publication. An additional label refining step using clustering is added in an attempt to improve the labels given by the basic method of labeling by price trends. The label refining method greatly improves the performance of the system. Marc et al.[2] objective is to forecasting intraday stock price trends with text mining techniques. They describe NewsCATS (News Categorization and Trading System), a system implemented to predict stock price trends for the time immediately after the publication of press releases. NewsCATS consists mainly of three components. The first component retrieves relevant information from press releases through the application of text preprocessing techniques. The second component sorts the press releases into predefined categories. Finally, appropriate trading strategies are derived by the third component by means of the earlier categorization. The findings indicate that a categorization of press releases is able to provide additional information that can be used to forecast stock price trends, but that an adequate trading strategy is essential for the results of the categorization to be fully exploited. ROBERT P. SCHUMAKER et al.[3] focuses on textual analysis of stock market prediction using breaking financial news. They examine a predictive machine learning approach for financial news articles analysis using several different textual representations: bag of words, noun phrases, and named entities. And investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. This estimate a discrete stock price twenty minutes after a news article was released.

Using a support vector machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables, the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price (MSE 0.04261), the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return).

Jheng-Long Wu et. al.[4] propose a stock price predication model which is combinational feature from technical analysis and sentiment analysis (SA). The features of sentiment analysis is based on a Pointwise mutual information (PMI) which is a term expansion method from multidimensional seed word. The features of technical analysis based on expert rule from trading information. Experimental results show that the use of sentiment analysis and technical analysis achieves higher performance than that without sentiment analysis in predicting stock price. Manisha V. Pinto et.al.[9] provide a framework for predicting stock magnitude and trend for making trading decisions by making use of a combination of Data Mining and Text Mining methods. The prediction model predicts the stock market closing price for a given trading day ‘D’, by analysing the information rich unstructured news articles along with the historical stock quotes. In particular, we investigate the immediate impact of the news articles on the time series based on Efficient Market Hypothesis (EMH). Key phrases provide semantic metadata that summarize and characterize documents. This framework incorporates Kea [10], an algorithm for automatically extracting key phrases from news articles. The prediction power of the Neural Network is used for predicting the closing price for a given trading day. The Neural Network is trained on the extracted key phrases and the stock quotes using the Back propagation Algorithm. Considering the news impact in analyzing the stock market behavior, leads to more precise predictions and as a result more profitable trades.

III. SYSTEM ARCHITECTURE

In this section, the architecture of NIPS (News Isolation and Prediction system) is described. This system consists of three modules: news collector, Isolator engine and comparator. Figure 1 shows the system architecture and interconnection between the modules. Descriptions of the individual modules are listed below:
1. News Collector

The news collector is used to collect news from different different resources. And analyzed by algorithm to extract key phrase. These phrases are act as an input to isolater engine.

2. Isolator Engine

The isolator engine sort the news articles coming from news collector, into predefined classes through If-Then rules of classification.

3. Comparator

The comparator is used to again further classified the earlier classified news. In this phase we use algorithm to further classified the result of isolator engine.

On arrival of a new article, the host application launches Extraction processing and the Isolater Engine in that order and generates appropriate rules depending on their outcome. Now historical daily prices and quotes of NSE is used from July 2011 to June 2013. The database of news article currently covers all news published by internet.

V. DATA ANALYSIS AND METHOD

For this study two types of data are required:-
1. News Articles.
2. Stock Prices.

A. Collection of News Articles

The news articles that are gathered will be labeled as positive, neutral or negative. The labeling techniques that are used to label the news articles are doing it by using stock prices. This means that the news articles that are to be gathered must be time stamped so they can be linked to stock prices. The web page Netfonds gathers time stamped news articles from various financial news sites and groups them together according to the companies they belong to. This means that in addition to fulfilling the requirement of having time stamped news articles[1].

B. Collection of Stock Prices

Daily Stock prices are collected about companies in National Stock Exchange (NSE) from July 2011 to June 2014. Daily quotes consists of date, time, open price, high price, low price, close price, and trading volume for that day.

C. How to label News Articles

After extraction of incoming news, we can classified incoming news into three predefined classes called positive, negative and neutral according to stock price. For labeling to news we use if-then rule based classification. If incoming news increase the stock price this news is labeled as positive. If decrease the stock price its label as negative. If coming news is not effecting the stock price this news is labeled as neutral. After categorization of news we will further classify the positive, negative and neutral news according to its type we can find it impact on stock price with the help of given below algorithm.

Input DATA // all types of news are collected for processing

Step 1: d1=initial value, dn=final value

And analyzed d1 to dn in system;

Step 2: if DATA=positive
Increase the index value
//Calculating index as sub categorized the news
Cond:1  If positive =finance
      More affect //index value is maximum
Cond:2  If positive =political
      Avg affected //index value is average
Cond:3  If positive =other
      Minimum affect
Step 3: if DATA=negative
      Decrease the index value
//Calculating index as sub categorized the news
Cond:1  If negative =finance
      More affect //index value is maximum
Cond:2  If negative =political
      Avg affected //index value is average
Cond:3  If negative =other
      Minimum affect
Step 4: if DATA=neutral
      Not affected to index value:
Step 5: //comparing the values
      If tested ≅ actual
      Count =results;
End

The above algorithm shows the effect of news in different perspective. In first step we check about the data ie in algorithm DATA which shows all the types of data in the form of financial news, these data are collected for preprocessing. In the next stage we define the data in there terms of initialization.

There are two terms use for initialization, first is d1 for initial value and another is dn for final value these both concern the limit by which data can be checked. After initialization we check the data in this limit.

The next step is comparison step here the data is compare as the positive or not, if the is positive then her it is further compared in three category like finance, political and general and there result of compare is also showing in this algorithm like if DATA is positive and financial then index value is maximum affected and if DATA is positive and political then it is affected but less. After this condition data is also compare in the terms of negative .If DATA is negative and financial then index value is maximum affected and if DATA is negative and political then it is affected but less. The third condition in both positive and negative is general which shows data lesser affection in the index point. Here the positive and negative data shows only the affection of index values that is how much the news affected to the index values. In next step there is third condition which compare DATA to neutral, if data is neutral then index value is not affected .Here neutral data shows the news which is not in the above category or other general news which is not concern to financial terms. At the last stage the final data is compared in the form of actual or if the data is actual the count the result.

D. Calculation

We create three categories of news articles: "positive", "negative", and "Neutral." In order to train the Isolator Engine with accurate examples for each category, we define as positive news all news articles that lead the stock price concerned to peak, with an increase of at least +3%, at some point during the 60 minutes immediately after publication and have an average price level in this period that is at least 1% above the price at the time of the release. The exact values are chosen arbitrarily, but their approximate levels are based on the following reflections. The first requirement is to identify those articles that have an immediate strong (positive) impact on the stock price, raising it by, say, +5% during the first 60 minutes. The second ensures that this effect does not hold only for a few trades, but that the press release provokes a shift of the average stock price by, say, +2% that persists for at least an hour after its publication. Sudden short-lasting price shocks, which can be caused, for example by interventions from market makers, can usually be observed during times of low activity or during pre- and post-market hours and should be eliminated from the training process. Since we operate with very short time intervals, the Beta of a stock and, mostly, the simultaneous fluctuations of the stock market (or an industry) are irrelevant.
On the other hand, all press releases leading to a maximum price drop of 5% and an average price level 2% below the price at the time of the release are considered negative news. This separation leads to classification of 347 news articles as positive news and 357, as negative news. The other 5,898 news articles are labeled "neutral." Several classifiers encounter problems when the categories in the training set vary significantly in frequency. In such cases there may be a bias towards prediction of the more common categories, leading to a worse category performance for the rarer categories [14]. To compensate for this peculiarity, we extract exactly 200 examples from each category and use these as training examples. The remaining examples are put into a holdout set that is later used to determine the model's accuracy. The 200 examples for each of the categories "positive News" and "negative News" are randomly extracted from the corresponding 347 and 357 news articles. Compared with systems implemented earlier, our approach is a novel one in that the 200 training examples for the category "Neutral" are randomly chosen from a subset only. This subset consists of those 1,166 (out of the 5,898) press releases that precede, simultaneously,

- The lowest maximum price change
- The highest number of price changes

of the corresponding stock during the 60 minutes following their publication. These restrictions make sure that the only press releases included are those that are not followed by large price changes or high volatility. In this way, we artificially create high selectivity between the three categories. The remaining 5,898 – 1,166 = 4,732 press releases in the "neutral" category are never used for training purposes. The preprocessing of the news articles proceeds as follows: During the feature extraction phase we use KEA [9] to extract key phrases. During feature selection we reduce each of the dictionaries to the 1,000 most meaningful terms.

In the context of classification tasks, the terms true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are used to compare the class labels assigned to documents by a classifier with the classes the items actually belongs to. True positives (TP) are examples that the classifier correctly labeled as belonging to the positive class of financial news. False positive (FP) are examples which were labeled by the classifier as belonging to the positive class having politics news. True negative (TN) are examples that the classifier correctly labeled as belonging to the negative class of political news.

At last there is false negative (FN) which are examples which were labeled by the classifier as belonging to the negative class having financial news. Table 1 shows how they belong together. Other evaluation measures like precision, recall and accuracy can easily be calculated from these four variables.

<table>
<thead>
<tr>
<th>Classified labels</th>
<th>Positive TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative FN</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

A common measure for classification performance is accuracy, or its complement error rate. Accuracy is the proportion of correctly classified examples to the total number of examples, while error rate uses incorrectly classified instead of correctly. In general, the better the classifier, the higher the accuracy is. However, one should be careful to use only accuracy when one is using skewed data. This is because when one class occurs significantly more than the other, the classifier might get higher accuracy by just labeling all examples as the dominant class then what it gets when it tries to classify some with the other class.

Precision and recall are two widely used metrics for evaluating performance in text mining, and in other text analysis field like information retrieval. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. In other words, a high precision means that the most of the documents labeled as positive are labeled correctly (but it might not have found many of the positive documents) and a high recall means that it has found most of the positive documents (but it might also have labeled many negative documents as positive). Accuracy, precision and recall are represented in tabular form.

<table>
<thead>
<tr>
<th>positive news</th>
<th>neutral news</th>
<th>negative news</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc</td>
<td>prec</td>
<td>rec</td>
</tr>
<tr>
<td>max</td>
<td>73%</td>
<td>7%</td>
</tr>
<tr>
<td>avg</td>
<td>68%</td>
<td>6%</td>
</tr>
<tr>
<td>min</td>
<td>71%</td>
<td>5%</td>
</tr>
</tbody>
</table>
VI. RESULT

Comparatively graphically representation of accuracy of news based stock prediction system is shown;

![Graph showing accuracy comparison](image)

VII. CONCLUSION

The news article based system NIPS(News Isolation and Prediction System)is able to do some good trading decisions beneficial for traders as a tool to help making better trading decisions. This system is differs from other system as if classifies the news again and again so that the impact of it can be further used by stock investors, so as to upgrade the gain and reduce the loss. Hence, this system is more beneficial as compared to previous systems.

REFERENCES
