Predicting Behaviors of Stock Market

Adhvik Shetty, Subham Chatterjee, Parimala R

Abstract*— Prices of stock depend on a variety of factors. Predicting and building a model is a daunting task to any analyst. To predict the behavior of stock market, one goes through the company news, economic and political news and global sentiments. Considering the large number of news articles, there are some which can be missed out. Also it is impossible to focus on each and every news article as soon as it is published on the internet. In this paper, we analyze the sentiment generated by news articles and correlate the sentiment with the actual change in stock market prices. This gives a deeper insight into the correlation and tells us how much news articles influence the stock market. After extensive research we have decided to use a hybrid technique involving machine learning and natural language processing concepts. We have used n–gram as the feature creation, chi square as the feature selection and support vector machines as the classification technique. Improving the accuracy of predicting stock market trends, we hope to aid investors in better decision making based on real time sentiment of news articles.

***Index Terms*— PRICES, CLASSIFICATION, ARTICLES, NEWS, POLITICAL**

I. INTRODUCTION

Stock market prices change every second. These changes are brought about due to variety of factors. Factors include sentiment regarding the company, political sentiments, economic sentiments, global sentiments, wars, disasters and many more. Investors across the globe scout for news articles which can indicate any of the above mentioned factors and then take a call whether to buy, sell or not get involved in the stock at all. There are at times, when new investors do not understand the sentiment thrown by a particular news article and end up taking the wrong decision while investing. This may end up being a costly mistake as sometimes lots of money are put up for the investment.

The number of news articles in the World Wide Web is infinite and sometimes investors miss out on important news regarding a company and that might end up being a costly mistake. Hence there is a need to automatic analysis of news articles which can extract the sentiment of a particular news article and assist investors to make a better decision. Such an automatic analyzer would have to run periodically based on the user's portfolio and be able to generate the sentiment [7]. In this paper, we have implemented a technique which generates the sentiment of the given news article and then correlates the sentiment with the actual change in stock price [5],[8].

Revised Version Manuscript Received on July 03, 2015.

Adhvik Shetty, PES Institute of Technology, BSK III Stage, Bangalore-560085, Karnataka, India.

Subham Chatterjee, PES Institute of Technology, BSK III Stage, Bangalore-560085, Karnataka, India.

Prof. Parimala R, Professor, Information Science Department, PES Institute of Technology, BSK III Stage, Bangalore-560085, Karnataka, India.

Many studies have taken place into this regard, but most of the prediction accuracy by analyzing news articles about corporate news of a company reaches 58%, which is not far from 50%, being the binary selection technique (positive or negative)[3],[11]. This does leave a lot of scope for improvement and provides a platform for more research in an area, which requires lot of careful technical analysis.

To analyze the news article automatically without any human help, we need to implement a machine learning technique. Machine learning techniques involve training a data set appropriate for the topic and classifying this training set into the classifiers. When we are implementing using actual news articles we then refer to this training set as a reference and then classify the news into positive, negative or neutral sentiment [13].

II. RELATED WORKS AND TECHNIQUES

Lot of research had been carried out in the past regarding textual analysis of news articles. Most of the techniques have followed the standard steps to achieve the results. Sentiment analysis typically involves the following steps or methods [4]:

1. Data Set
2. Feature Processing
3. Machine Learning

Data set involves the set of texts which will be taken as the reference while undergoing sentiment analysis [6]. It is the raw data taken from a relevant source on which feature processing takes place.

Feature processing is a very important step to get good results in classification. It involves the set of keywords which are retained from the data set which matter a lot to the classifier. The first step in feature processing is feature extraction. This is generally carried out by cleaning the data set. Cleaning involves removing extra white spaces, removing symbols like #, %, @ etc. [2] After cleaning the data set, extraction of words take place from the data set. [10] The range of possible features include single discrete words, n-grams, noun phrases, statistical information about the words.

1. Single Discrete words – These include each and every word considered as a single entity.
2. N –gram – These include phrases which include n words at a time. (A bird in the sky becomes 'A bird', 'bird in', 'in the', 'the sky', when n equals 2)
3. Single noun phrases – A noun accompanied by adjectives belong to this category. (A blue bird)
4. Statistical information – This includes number, position of words.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

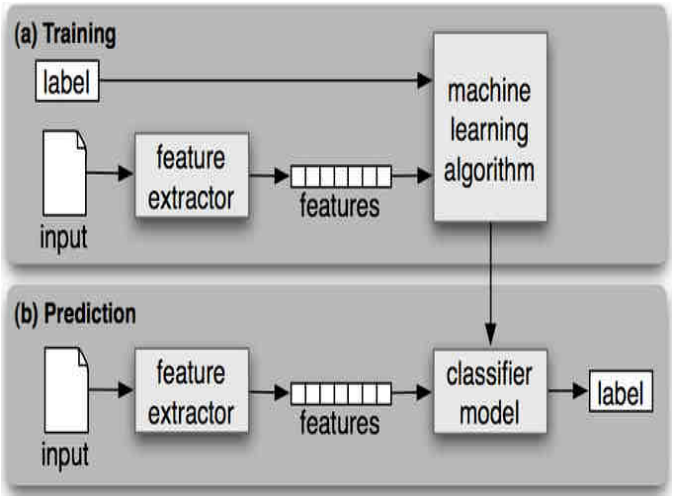

Fig. 2.1 Steps in the process [9]

The next step include feature selection. This method includes selecting a subset of the feature extracted. This is done by analyzing what is best from the available feature that will yield the best result during the classification [1]. This can be done using the following ways:

1. Manual Identification – Domain experts manually identify which words are the most relevant words based on the type of classification that will be done. This is a cumbersome process and takes a lot of time.
2. Feature selection based on endogenous information – In this case, the feature is selected based on information in the dataset.
3. Feature selection based on exogenous information – Market feedback is used in this case to select the most relevant words that should be added to the feature vector.

Feature representation is the way the feature vector is stored in a computer or processed by the classification algorithm. There are many ways to represent the feature, but mostly standard data structures are used to represent them.

The third step includes the machine learning path. This is the component which applies the learning from the training data set to apply to the test data. There are many approaches to this method like SVM, Neural Network, Maximum Entropy but all these methods use the same concept – apply learning from training set to test set. These approaches are used to classify the test set into one of the classification that are part of the training set. Previous works have shown that it does not matter to an extent what classifier one uses, as the most important step of the process is the feature processing. Only if the training data is accurate, the test data will follow the same pattern. After all the machine learning algorithms try to find patterns in the training set that also exist in the test set.

SVM constructs a hyperplane that divides the plane into two regions. Each data is mapped onto the plane, and SVM helps it to classify into the particular category based on it [12].

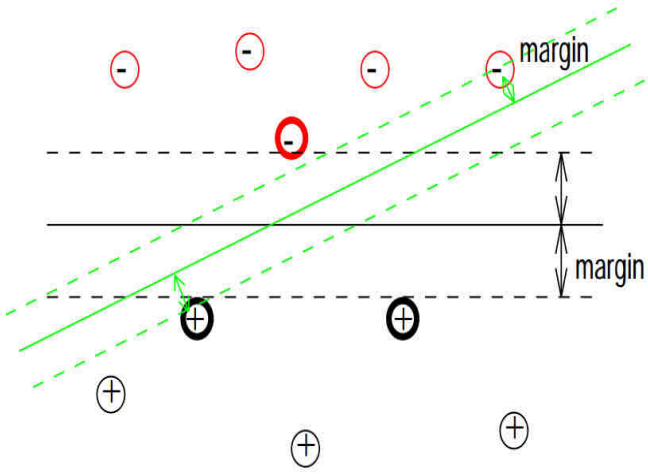

Fig. 2.2 Representation of SVM [12]

III. METHODOLOGY

To analyze the sentiment, the first step includes gathering a data set. In our paper, we have considered the Cornell data set [needs changing]. This data set contains articles which has been categorized as positive or negative.

After the data set is acquired, the most important step of the process awaits. It is the feature vector creation process. The feature creation process has been divided into four parts. The first part involves cleaning of the data. This involves removing extra white spaces, removing punctuation marks, removing symbols like #, $, %, ^ etc. This ensures that the data is devoid of any irregularities. We then proceed to remove stop words. These words are generally the prepositions like ‘at’, ‘on’, ‘for’, ‘which’ and many more. These words generally do not contain any meaning and are therefore not useful in any way to our prediction. There might also be cases where in these words can alter the prediction and end up giving incorrect predictions. Hence, we get rid of such words before constructing feature vector [1],[2].

The next step in the process is the feature representation. After preprocessing and removal of stop words, we now have words which will be used to put in the feature vector. In our methodology, we have used n –grams to represent the feature set. N- grams consists of n words appearing together in the feature set. N –grams can be of many types such as unigrams, bigrams, trigrams etc. [5] Hence we transform the word list into an n – gram format and represent the feature list as list of n words in one line.

Once, we have acquired the feature list, we need to make sure that the most appropriate words are present in the feature list. This is to make sure that as we are taking raw data from an external source, there are cases when words from other classifier gets mixed with each other. Hence, we need to select only a subset of the feature list that consist of the words which are in general have the highest frequency. In our process, we have used Chi-Square method to select the feature list that will be used in the classification.

Chi square is a test of measuring the divergence from the distributions. It is one of the most accurate methods of measurement though there are instances where it is known to behave erratically when the data size is very small [4]. However in our case the data size being quite large would eventually give pretty accurate results.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Chi Chi-Squared[4]

$$t(tp,(tp+fp)P_{pos}) + t(fn,(fn+tn)P_{pos}) + t(fp,(tp+fp)P_{neg}) + t(tn,(fn+tn)P_{neg})$$

where $t(count,expect) = (count - expect)^2 / expect$

Notation:

tp: true positives = number of positive cases containing word	*fn*: false negatives
fp: false positives = number of negative cases containing word	*tn*: true negatives
pos: number of positive cases = *tp* + *fn*	$P_{pos} = pos / all$
neg: number of negative cases = *fp* + *tn*	$P_{neg} = neg / all$
tpr: sample true positive rate = *tp* / *pos*	$P_{word} = (tp+fp) / all$
fpr: sample false positive rate = *fp* / *neg*	P_{word} = 1-P(word)
precision = *tp* / (*tp*+*fp*)	*recall* = *tpr*

Fig. 3.1 Chi- square formula

For each and every word in the feature list we calculate the magnitude by applying the chi- square formula. We then sort the list and select only the top 50 % as it is an indication that these words are significantly important in the classification and depicts the accurate representation the in the feature list [1].

The next and the final step of the process is the classification. Classification helps in putting the test cases into the right group that is positive or negative. It provides the prediction after analyzing the test sentence. In our study, we have implemented the LIBSVM library to proceed with the classification. LIBSVM acts a library for Support Vector Machines. It can be found at http://www.csie.ntu.edu.tw/~cjlin/libsvm [12]. It provides us with an interface which consists of methods like saving the trained model, loading the trained model, training the classifier etc.

Once we provide the input which consists of news from the financial world, LIBSVM predicts the classification by loading the saved model and analyzing based on it. It provides us with the accuracy, prediction, magnitude among others. Based on this classification, we compare with the change in the actual stock price and provide the results.

IV. RESULTS

All the 30 stocks in the BSE SENSEX were taken into consideration and they were run against our algorithm.
We achieved a total of 66.67 % of accuracy when we take change in stock price in correlation with the prediction. The closest one has come up to so far has been 65%. Stock prediction can turn out to be low in terms of accuracy as prices depend on a wide variety of factors and not on the news alone. This includes wars, natural calamities, political sentiments, internal disturbances and also news regarding the connected sectors.
In a pilot study we measured the accuracy unigram, bigram, trigram, 4-gram and 5-gram. We found the trigram model to give the best accuracy. Also pre processing the data set gave a 10.2% increase in accuracy.

Stock Name	Our Prediction	Magnitude	Actual change is stock price	Correlation
Axis Bank	Up	3.32	Negative	No match
Bajaj Auto	Up	3.44	Negative	No match
Bharti Airtel	Down	-0.73	Negative	Match
BHEL	Down	-1.93	Negative	Match
CIPLA	Down	-1.64	Negative	Match
Coal India	Up	5.63	Positive	Match
Dr. Reddy	Down	-3.24	Negative	Match
GAIL	Up	2.95	Positive	No Match
HDFC	Down	-2.66	Negative	Match
HDFC Bank	Down	-2.66	Negative	Match
Hero Motor Corp	Up	2.87	Negative	No Match
HindalCo	Down	-1.55	Negative	Match
Hindustan Unilever	Down	-0.36	Negative	Match
ICICI Bank	Up	0.9722	Negative	No Match
Infosys	Down	-0.94	Negative	Match
ITC	Up	2.66	Positive	Match
Larsen and Toubro	Up	3.77	Negative	No Match
Mahindra	Down	-2.03	Negative	Match
Maruti	Up	1.166	Positive	Match
NTPC	Up	2.76	Negative	No Match
ONGC	Up	3.10	Negative	No Match
Reliance	Down	-0.31	Negative	Match
SBI	Up	1.46	Positive	No Match
SSLT	Up	1.67	Positive	Match
Sun Pharma	Down	-1.62	Negative	Match
Tata Motors	Down	-1.89	Negative	Match
Tata Power	Down	-1.06	Positive	No Match
Tata Steel	Up	0.57	Positive	Match
TCS	Up	1.77	Positive	Match
Wipro	Down	-1.10	Negative	Match

V. CONCLUSION

Predicting stock prices by analyzing news articles is quite challenging and prone to errors as there involves lot of complexity based on which stock prices change. Also, sentiment analysis is error prone as the results depend a lot on the training data set and the feature vector. An attempt has been made in this direction and sentiment analysis of news articles have been done, based on which behaviors of stock market has been predicted. Better feature creation and feature selection can be used to enhance the feature vector which in turn will yield better results. Preprocessing and cleaning of data also helped improve results. One of the most efficient classification technique, Support Vector Machines has been used to classify the news as positive or negative. The results are found satisfactory for the techniques and algorithms used. This application can be used by equity market analysts, stock brokers and anyone wishing to enter the stock market.

ACKNOWLEDGMENT

At the outset, we would like to thank our college, PESIT, Bangalore for an opportunity to carry out a research on one

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

of the most important and ongoing topics in the world. We also thank our HOD, Dr. Shylaja S.S, Prof. Parlimala and Dr. Natrajan who helped and motivated us in all endeavors.

REFERENCES

1. An extensive empirical study of feature selection metrics for text classification by George Gorman, Journal of Machine Learning Research 3(2003)
2. Preprocessing the Informal Text for efficient Sentiment Analysis by I Hemalatha, Dr. GP Saradhi Verma, and Dr A Govardhan, Internal Journal of Emerging Trends and Technology in Computer Science
3. Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang and Xiaotie Deng. “News impact on stock price return via sentiment analysis.” Knowledge-Based Systems(2014).
4. Gonçalves, Pollyanna, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. "Comparing and Combining Sentiment Analysis Methods." ACM(2013).
5. Butler, M., Keselj, V. 2009. “Financial Forecasting using Character N-Gram Analysis and Readability Scores of Annual Reports”, Advances in AI
6. Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. In Proceeding of the intl. conference on recent advances in natural language processing. Borovets, BG.
7. S. Das and M. Chen. Yahoo! for anazon: Extracting market sentiment from stock message boards. In Proc. of the 8th APFA, 2001.
8. M. Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. http://sentistrength.wlv.ac.uk/documentation/ SentiStrengthChapter.pdf
9. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In ACL Conference on Empirical Methods in Natural Language Processing, pages 79–86, 2002.
10. Antweiler, W., Frank, M.Z. 2004. "Is all that talk just noise? The information content of internet stock message boards”, The Journal of Finance, Volume 59, Number 3, June 2004, pp. 1259-1294
11. Gidofalvi, G. & Elkan, C. 2003. “Using News Articles to Predict Stock Price Movements. Technical Report”, Department of Computer Science and Engineering, University of California, San Diego
12. Joachims, T., 1998. “Text categorization with support vector machines: Learning with many relevant features”, Proceedings of the European Conference on Machine Learning, Springer-Verlag.
13. Schumaker, R.P., Chen, H. “Textual analysis of stock market prediction using breaking financial news: the AZFin Text System”, ACM Transactions on Information Systems 27 (2) (2009).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.