International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(1), pp.071-077
DOI: http://dx.doi.org/10.21172/1.81.009
e-ISSN:2278-621X

PREDICTIVE SENTIMENTAL ANALYSIS OF STOCK TWEETS

Rajesh K Ahir[1] and Mital B Ahir[2]

Abstract- Now-a-days social networking sites are at the boom, so large amount of data is generated. Millions of people are sharing their views daily on micro blogging sites, since it contains short and simple expressions. In this paper, we will discuss about a paradigm to extract the sentiment from a famous micro blogging service, Twitter, where users post their opinions for everything. Firstly we will convert unformatted data into formatted data in preprocessing level. Secondly we will use a classification algorithm for classifying tweets as positive or negative. Based on this approach the output will be in the form of graph representing the prediction of stock tweets.
Keywords – Sentimental Analysis, Classifiers, Pre-processing, Text Mining

I. INTRODUCTION

Over the past few years, an increasing number of people have begun to express their opinion through social networks and micro blogging services. Twitter, as one of the most popular of these social networks, has become a major platform for social communication, allowing its users to send and read short messages called 'tweets'.

Twitter allows people to share their opinions and thoughts openly about every topic, discussion point or product in which they are interested in sharing their opinions about. Therefore Twitter is a good platform to search for potentially interesting trends regarding noticeable topics in the news or popular culture. Twitter is much casual and less reliable in terms of language. Users cover a huge collection of topics which attracts them and use many signs such as emoticons to express their views on many facets of their life[11].

Among the different software that can be used to analyze twitter, R offers a wide variety of options to do lots of interesting and fun things. In this project we have used RStudio as its pretty much easier working with scripts as compared to R. Sentiment analysis is defined as the use of natural language processing, text mining and computational semantics to identify and extract particular information in source material. It provides a means of tracing opinions and views on the web and determines if they are positively or negatively received by the public. Its purpose is to process unstructured information and to extract expressive numeric catalogues from the text, allowing the application of various data mining algorithms to explain the textual dataset.

[1] *G H Patel College of Engineering & Technology, Anand, Gujarat, India 388120*
[2] *Student, IIET, Dharmaj, Anand, Gujarat, India 388120*

The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, or you could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In the most general terms, text mining turns "text into number" which can then be incorporated in other analyses.

Applications of text Mining are analyzing open ended survey responses, automatic processing of messages, emails, etc., analyzing warranty or insurance claims, diagnostic interviews, etc., investigating competitors by crawling their web sites.

II. LITERATURE REVIEW

Over the past many years, important changes have taken place in the environment of financial markets[11]. The growth of dominant communication and trading facilities has inflated the scope of selection for investors. Predicting stock return is an important task. These are some of the papers which we have referred for our Project:

[6] has made a study on building a stock buying/selling alert machine using neural network techniques. The results generated on the basis of past data were proved to be 74% accurate.

[4] presented a framework based on text mining to define the sentiment of news articles and show its impact on energy demand.

[6]presented an approach which consisted of associating entities with sentiments & aggregating & scoring each entity.

[7,9] provided an overview of application of data mining techniques such as decision tree, neural network, association rules, and factor analysis and in stock markets.

[8] proposed a prediction stock price or financial markets has been one of the biggest challenges to the AI community. Various technical, fundamental, and statistical indicators have been proposed and used with varying results.

[12]surveyed some recent literature in the domain of machine learning techniques and artificial intelligence used to predict stock market movements.

[5] proposed a new approach for fast forecasting of stock market prices. The proposed approach uses new high speed time delay neural networks.

[1] surveyed a concept of different data mining algorithms for cart item of shopping.

III. CLASSIFICATION ALGORITHMS

The Financial subject that has attracted researchers' attention for many years. It includes a supposition that important information publicly available in the past has some predictive associations to the future stock returns. There are three different Classification algorithms who achieved great success for text categorization which are as follows:

A. Naive Bayes

It is a classification technique based on Bayes' theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).
Look at the equation below:

$$P(c|x) = P(x|c)\, P(c) / P(x) \quad (1)$$

$$P(c|x) = P(x_1|c) * P(x_2|c) *.....* P(x_n|c) * P(c) \quad (2)$$

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes). P(c) is the prior probability of class. P(x|c) is the likelihood which is the probability of predictor given class. P(x) is the prior probability of predictor.
Advantages:
1. Fast to train and fast to classify.
2. Not sensitive to irrelevant features.
3. Handles real and discrete data.
4. Handles streaming data as well.
Disadvantages:
1. Very simple representation doesn't allow for rich hypotheses.

B. kNN Classifier

The kNN (k-Nearest Neighbour) algorithm is a non-parametric algorithm [2] that can be used for either classification or regression. Non-parametric means that it makes no assumption about the underlying data or its distribution. It is one of the simplest Machine Learning algorithms, and has applications in a variety of fields, ranging from the healthcare industry, to the finance industry. For each data point, the algorithm finds the k closest observations, and then classifies the data point to the majority. Usually, the k closest observations are defined as the ones with the smallest

Euclidean distance to the data point under consideration. For example, if k = 3, and the three nearest observations to a specific data point belong to the classes A, B, and A respectively, the algorithm will classify the data point into class A. If k is even, there might be ties. To avoid this, usually weights are given to the observations, so that nearer observations are more influential in determining which class the data point belongs to. An example of this system is giving a weight of 1/d to each of the observations, where d is distance to the data point. If there is still a tie, then the class is chosen randomly.
Advantages:
1. Robust to noisy training data.
2. Effective if training data is large.
Disadvantages:
1. Computation cost is quite high.
2. Distance based learning is not clear as in which type of distance to use and which attribute to use to produce best results.

C. Decision Tree

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. Some of the well-known decision tree algorithms are ID3, C4.5 and CART.
ID3 algorithm is an expansion of concept learning theory by [10]. It is a recursive procedure using divide and conquer approach, which supports only nominal attributes. Information gain is used to select an attribute to split. It does not give accurate result when there is too-much noise or details in the training data set, thus a an intensive pre-
processing of data is carried out before building a decision tree model with ID3. C4.5 is developed by [10], uses gain ratio for selection of attribute for splitting. It provides an improvement over ID3 as it deals with nominal and numerical attributes as well as able to handle missing and noisy data. Pruning in C4.5 takes place by replacing the internal node with a leaf node thereby reducing the error rate. Classifier generated by C4.5 can be expressed not only in terms of decision tree but also in more comprehensible rule set form. The major disadvantage of rule set form is that it require large amount of CPU time and memory.
CART (Classification and Regression Trees) proposed by [10], uses Gini index measure for selecting attribute for splitting. Test in CART is always binary. CART prunes trees using a cost complexity model whose parameters are estimated by cross-validation. The advantages and disadvantages of Decision Tree Induction are as follows.
Advantages:
1. Decision Trees are very flexible, easy to understand and easy to debug.
Disadvantages:
1. Simple decision trees tend to over fit the training data more so you need to process it more.

IV. PROPOSED SYSTEM

Sentiments are the words or sentences that represent view or opinion that is held or ex- pressed that can be positive or negative. We are going to use RStudio for our project as it is the most

efficient platform to deal with statistical data. RStudio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

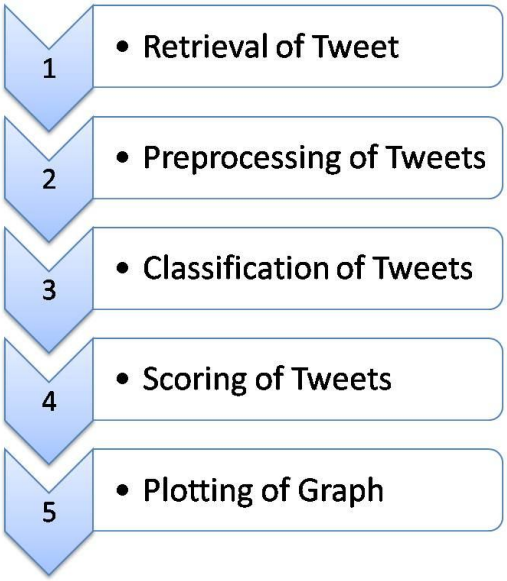


Figure 1. Proposed Methodology

After extracting tweets, we will also consider features like emoticons, neutralization, negation handling and capitalization as they have recently become a huge part of the internet language. The proposed Sentiment Analysis on twitter data is based on some important parts viz Data Extraction, preprocessing of extracted data and classification.

The following steps will expound the process of the proposed system:

1. Retrieval of stock tweets

It will include creating a twitter application, installing and loading R packages, creating and storing twitter authenticated credential objects and extracting stock tweets.

2. Pre-processing of extracted tweets

Extracted tweets are in an unformatted form. They cannot be used for analysis purpose. So, they need to be processed for obtaining better results. We extract tweets i.e. short messages from twitter which are used as raw data. This raw data needs to be preprocessed. So, preprocessing involves following steps:

i) Filtering:

Filtering is nothing but cleaning of raw data. In this step, URL links http://twitter.com, special words in twitter (e.g. RT which means ReTweet), user names in twitter (e.g. @Ron - @ symbol indicating a user name), emoticons, digits are removed.

ii) Tokenization:

Tokenization is nothing but Segmentation of sentences. In this step, we will tokenize or segment text with the help of splitting text by spaces and punctuation marks to form container of words.

iii) Removal of Stop words:

Articles such as a, an, the and other stop words such as to, of, is, are, this, for removed in this step.

3. Classification of tweets

After processing the extracted tweets, we need to classify them into positive or negative tweets. For classification many algorithms can be used as shown in Section 3. We will be using Nave-Bayes Algorithm for classification purpose. We will be comparing the extracted tokens with the available sentiment word list in R. The word will be match with text mining library called tm.plugin.tags.

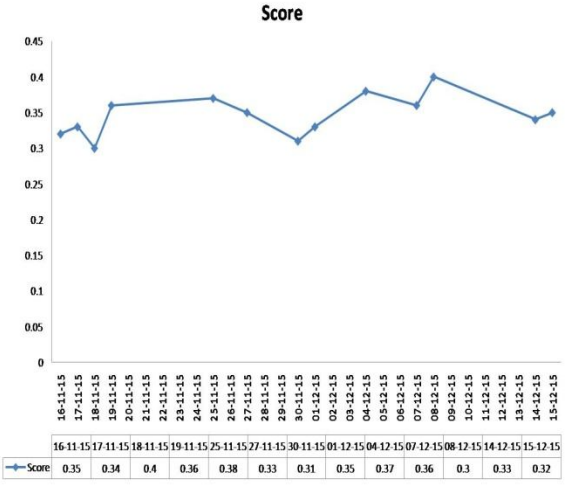
4. Scoring classified tweets

An instance is classified as positive if the count of positive words is greater than or equal to the count of negative words[12]. Similarly, an instance is negative if the count of negative words is greater than the count of positive words.

$S_I = \text{sign}(N_p - N_n)$,where S_I is the score of the instance, N_p is the number of positive words, N_n is the number of negative words and the sign functions returns the sign of its argument.

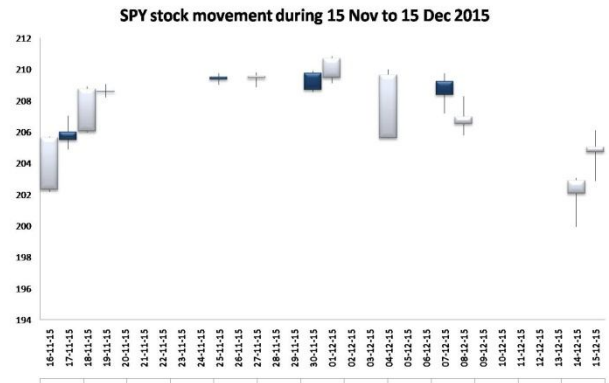
5. Plotting Graph.

Below figures shows a analysis of the score obtained by considering filtered tweets to the actual stock price variation for SPY stock between 15 Nov and Dec 15, 2015.



	16-11-15	17-11-15	18-11-15	19-11-15	25-11-15	27-11-15	30-11-15	01-12-15	04-12-15	07-12-15	08-12-15	14-12-15	15-12-15
Score	0.35	0.34	0.4	0.36	0.38	0.33	0.31	0.35	0.37	0.36	0.3	0.33	0.32

Figure 2 (a). Movement based on score



	16-11-15	17-11-15	18-11-15	19-11-15	25-11-15	27-11-15	30-11-15	01-12-15	04-12-15	07-12-15	08-12-15	14-12-15	15-12-15
Open	204.7	202.0700	206.4900	209.23	205.61	209.44	209.75	209.4299	209.5	208.59	206.0399	205.9900	202.3200
High	206.11	203.05	208.2899	209.73	209.97	210.8200	209.89	209.8	209.7400	209.05	208.8999	207.0399	205.69
Low	202.87	199.95	205.78	207.2	205.61	209.11	208.56	208.86	209.01	208.2	205.9900	204.8800	202.1799
Close	205.03	202.8999	206.95	208.3500	209.62	210.6799	208.69	209.56	209.3200	208.55	208.73	205.47	205.62

Figure 2 (b). SPY Actual stock movement

From above figure it can be easily observed that actual movement of stock and score from the tweets are sparks a significance impact of prediction.

V. CONCLUSION

Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. In the survey, we found that social media related features can be used to predict sentiment in Twitter. We will use classification algorithms in this project for classifying sentiment of tweets extracted from twitter which can be positive or negative. This tool can be helpful for the investors to take right decision regarding their stocks based on the analysis of the historical prices of Stocks in order to extract any predictive information from the historical current data.

VI. FUTURE WORK

Our method gives appropriate results not more accurate ones. So for accurate results more relevant words related to stock market should be taken in to account, analyzed and processed.

Another problem is a tweet like I am not happy is a negative tweet, but with the help of our proposed method the only feature extracted would be happy which is a positive word and therefore the tweet will be classified as a positive tweet, but in reality the tweet is a negative. So further work can be done to solve this problem.

SUPPLEMENTAL MATERIALS

Title: Dataset.

SPY Stock data set: Data set used in the implementation of system in section 4. (.csv file)

REFERENCES

[1] Rajesh K Ahir et al. Algorithem for mining frequent patterns. International Journal of Advance research on Computer science and control Engineering, 2(12), December 2013.

[2] Khalid Alkhatib et al. Stock price prediction using k-nearest neighbor. Internation Journal of Business Humanities and Technology, 3(3), March 2013.

[3] E.F.Fama and K.R.French. The crosssection of expected stock returns. The Journal of Finance, 47:427–465, 1992.

[4] E.F.Fama and K.R.French. Common risk factors in the returns on stocks and bonds. The Journal of Finance, 33:3–56, 2010.

[5] E.Hajizadeh, H. Ardakani, and J.Shahrabi. Application of data mining techniques in stock marktes:a survey. Journal of Economics and International Finance, 2(7):109– 118, July 2013.

[6] N. Godbole, M.Srinivasaih, and S.Skiena. Large-scale sentiment analysis for news and blogs. In Proceedings of International Conference on Weblogs and Social Me- dia(ICWSM), 2007.

[7] J.Bean. R by example:mining twitter for consumer attitudes towards airlines, 2011. Boston Predictive Analytics Meetup Presentation.

[8] M.Hazem, El-Bakry, and Wael A.Awad. Fast forcasting of stock market prices by using new high speed time delay neural networks. International Journal of Computer and Information Engineering, 4(2):138–144, 2010.

[9] Cao Q et al. A comparison between fama and french's model and artificial neural networks in predicting the chinese stock market. Computers & Operation Research, 32:2499–2512, 2012.

[10] Qasem et al. Predicting stock prices using data mining techniques. The International Arab Conference on Information Technology(ACIT 2013), 2013.

[11] R.Goonatilake and S.Herath. The volatility of the stock market and news. Internation Research Journal of Finance and Economics, 11:53–65, 2007.

[12] S.Soni. Application of anns in stock market prediction:a survey. International Journal of Computer Science & Engineering Technology(IJCSET), 2(3):71–83, 2011.

[13] S.Theussl. tm.plugin.tags:text mining plugin:tag categories, 2010. R package.

[14] S.Theussl, I.Feinerer, and K. Hornik. Distributed text mining with tm. In In Proceedings of R Finance, 2010.