IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Predicting Stock Market Investment Using Sentiment Analysis

Shantanu Pacharkar[1], Pavan Kulkarni[2], Yash Mishra[3], Amol Jagadambe[4], S.G.Shaikh[5]

Student, Computer Engineering, Sinhgad Institute of Technology, Lonavala, India [1,2,3,4]

Professor, Computer Engineering, Sinhgad Institute of Technology, Lonavala, India [5]

Abstract: When a customer or a new comer invests into a stock market, they want to attain higher profits in short period of time. With less amount of knowledge, they have. The process of attaining higher profit gets very difficult. Sometimes this situation often creates more losses to customers rather than profit. Out of all the books offering investing advice to research papers analysing mathematical prediction models, the stock market has always been centre of attraction for public and academic interest. Number of publications propose strategies with good profits, while others demonstrate the random and unpredictable behaviour of share prices. Shared prices aren't really based on how a company works. It's based on how mass psychology works. Sentiment Analysis is one of the most popular technique which is widely been used in every industry. Extraction of sentiments from user's comments is used in detecting the user view for a particular company. Sentiment Analysis can help in predicting the mood of people which affects the stock prices and thus can help in prediction of actual prices. Stock market prediction is the act of trying to determine the future value or other financial instrument traded on a financial exchange. In this paper, project is overall based upon the myriad data which is going to be mined from various stock related portals, social media, etc. and after fetching the desired data they have been used for the predictions of related results using NB classifier (Naive Bayes classifier).

Keywords: Sentiment Analysis, Stock Market, Public review, Naive Bayes classifier.

I. INTRODUCTION

If there is a company having product 'P' and has a great demand in market and there is a hike for that product. But there is a limit to what that company can produce that production. There can be various reasons for the limitation of the production. One of the main reason is the requirement of more capital for the production. In that case company keeps some percent (at least 50%) of shares in the market and sells it. A 'Share' is one of the equal parts into which a company's capital is divided, entitling the holder to a proportion of the profits. A share is also known as stock. The stock of a corporation is constituted of the equity stock of its owners. A single share of the stock represents fractional ownership of the corporation in proportion to the total number of shares. Now here, in order to buy that share or stock of that company, the customer or in more corporate language the investor have to give some money to the company in order to increase its production and in return that investor becomes an equal shareholder of that company. After that whenever there is a profit in company the share prices will go up and then the investor can sell the shares with nice profit.

But the main question arises that when there will be a hike in the price of that share/stock that the investor had already bought. In this way the investor can lose a lot of money if they invested in a company's shares blindly. In order to avoid that, the investors use various methods to predict the prices of shares and market and then invest accordingly. Various algorithms or programs are used.

Sentimental Analysis also known as Opinion Mining is an area that uses Natural Language Processing and Text Analysis that helps in building a system that identifies and extract information in source material. An initial task in sentimental analysis is to determine the polarity of a specified text at the document level, sentence level or aspect level. In the finance field, stock market and its trends are extremely volatile in nature. It attracts researchers to capture the volatility and predicting its next moves. Investors and market analysts study the market behaviour and plan their buy or sell strategies accordingly. As stock market produces large amount of data every day, it is very difficult for an individual to consider all the current and past information for predicting future trend of a stock.
Earlier studies on stock market prediction are based on the historical stock prices. Later studies have debunked the approach of predicting stock market movements using historical prices. Stock market prices are largely fluctuating. The efficient market hypothesis (EMH) states that financial market movements depend on news, current events and product releases and all these factors will have a significant impact on a company's stock value [2]. Because of the lying unpredictability in news and current events, stock market prices follow a random walk pattern and cannot be predicted with more than 50% accuracy [1].

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

II. LITERATURE SURVEY

The most interesting task is to predict the market. So many methods are used for completing this task. Methods, vary from very informal ways to many formal ways a lot. These technologies are categorized as Prediction Methods, Traditional Time Series, Tech Analysis Methods, Mach Learning Methods and Fundamental Analysis Methods. The criteria to this category is the kind of tool and the kind of data that these methods are consuming in order to predict the market. What is mutual to the technique is that they are predicting and hence helping from the market's future behaviour.

Stock price trend prediction is an active research area; as more accurate predictions are directly related to more returns in stocks. Therefore, in recent years, significant efforts have been put into developing models that can predict for future trend of a specific stock or overall market. Most of the existing techniques make use of the technical indicators. Some of the researchers showed that there is a strong relationship between news article about a company and its stock prices fluctuations. Following is discussion on previous research on sentiment analysis of text data and different classification techniques.

Nagar and Hahsler in their research [3] presented an automated text mining based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analysed using Natural Language Processing (NLP) techniques. A sentiment metric, called News Sentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. They have used various open source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of News Sentiment shows a very strong correlation with the actual stock price movement.

Yu et al [4] present a text mining based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified and then presented as a time series and compared with fluctuations in energy demand and prices.

J. Bean [5] uses keyword tagging on Twitter feeds about airlines satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings. We have used the sentiment detection algorithm based on this research.

A growing amount of literature is devoted to developing new tools and models for sentimental analysis. Previous studies had concentrated on a large group of population using social information in prediction of consumer's attitude towards a company [4]. Numerous studies showed that using a mix approach can improve classification scheme [5]. The most common use of sentimental analysis is analysing of twitter tweets and demonstrating the top trends in marketplace [6] Sentimental analysis is also used in sales forecast of a product by examining tweets and posts from face book [7]. Sentimental analysis had been conducted on stock linked tweets which were collected for a period of 6-month [8]. In order to reduce noise, selection of tweets containing tags of top 100 companies was considered. Each tweet was classified using a Naive Bayes method and a set of 2,500 tweets were trained. Results displayed that sentiment indicators are related with unusual returns and stock volume is linked with trading volume [9].

Sentimental Analysis was applied to tweets extracted from Twitter and news headlines to generate new predictors for investment [10]. From the collected data, they choose a random sample and defined each tweet as bullish or bearish if it contains those terms. They displayed that Twitter sentiment indicator and the occurrence of monetary terms on Twitter are statistically significant predictors of regular market returns. Sentimental analysis was also performed on a micro blogging service entirely devoted to stock market [11] They collected 62,100 blog posts from stocktwits.com, for a time span of three months. The sentiment of the posts was classified using a machine learning algorithm known as NB classifier to generate a learning model. They proved that the mined sentiment have strong analytical value for coming market directions. Sentimental Analysis was used to forecast the closing index of Tata Services and an accuracy of 85.99\% was found in the process [13].

Sentimental analysis is often used to build a social behaviour graph on human's online behaviour to find the correlation between trading and volume prices of stocks [14]. Sentimental Analysis was also performed on the data extracted from SentiWordNet using a hybrid selection model to show how market trends affects a product popularity and rate [15].

III. PROPOSED MODEL

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry","sad", and "happy".

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Precursors to sentimental analysis include the General Inquirer, which provided hints toward quantifying patterns in text and, separately, psychological research that examined a person's psychological state based on analysis of their verbal behaviour.

Figure 1: Mood-Meter

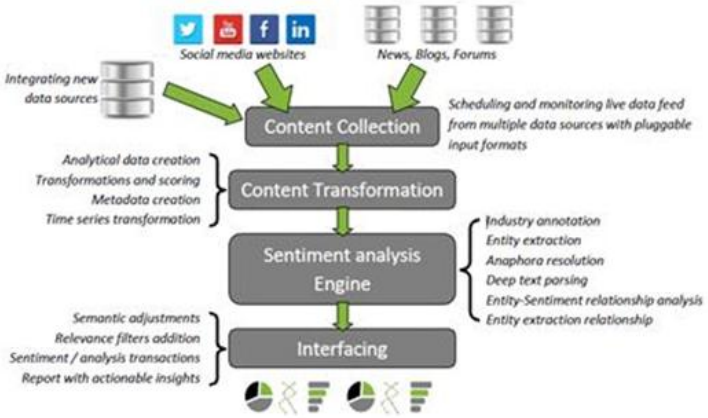
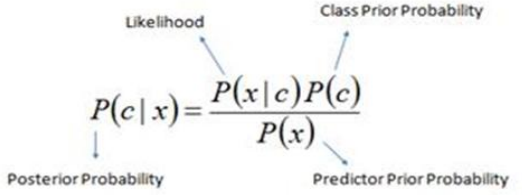


Figure 2: Sentiment Method

It is a classification technique based on Bayes theorem with an assumption of independence between 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, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.

Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive 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: Bayes rule



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Here,
P(c|x) is the posterior probability of class (target) given predictor (attribute). 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.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

A. Training the Naive Bayes Classifier

Well simply use the frequencies in the data. For the document prior P(c) we ask what percentage of the documents in our training set are in each class c. Let Nc be the number of documents in our training data with class c and Ndoc be the total number of documents. Then:

$$P(c) = \frac{Nc}{Ndoc}$$

To learn the probability $P(f_i|c)$, well assume a feature is just the existence of a word in the documents bag of words, and so well want $P(w_i|c)$, which we compute as the fraction of times the word w_i appears among all words in all documents of topic c. We first concatenate all documents with category c into one big category c text. Then we use the frequency of w_i in this concatenated document to give a maximum likelihood estimate of the probability:

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)}$$

Here the vocabulary V consists of the union of all the word types in all classes, not just the words in one class c. There is a problem, however, with maximum likelihood training. Imagine we are trying to estimate the likelihood of the word fantastic given class positive, but suppose there are no training documents that both contain the word fantastic and are classified as positive. Perhaps the word fantastic happens to occur (sarcastically?) in the class negative. In such a case the probability for this feature will be zero:

$$\hat{P}(\text{“fantastic”}|\text{positive}) = \frac{count(\text{“fantastic”}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

But since naive Bayes naively multiplies all the feature likelihoods together, zero probabilities in the likelihood term for any class will cause the probability of the class to be zero, no matter the other evidence!

IV. WORKING

A. Basic Overview:

This project is basically a webapp based on two sides, i.e. frontend and backend. Front end is the common webpage which can be accessed from any browser and backend is the python programming which will be running on server side. As the user/ investor will search the company name or product name, the recent comments and reviews regarding that company will be fetched in runtime and then sentiment analysis will be performed on that data. By this, we will get the analysed review of the product or company’s performance.

B. Technology Used:

- Python v3.6
- IDE (Pycharm)
- Django Framework
- Various APIs (Yahoo Finance API, The Guardian API, etc.)
- Visual Studio / Atom code editor (UI Designing)
- SQLite (Database)

C. Modules:

- News Collection,
- Text Pre-processing,
- Polarity Detection,
- Document Representation,
- Classifier Learning,
- System Evaluation,
- Test The Model with New Data,
- Plot Scoring of News Sentiment,
- Observe The Relation Between News/Tweet Sentiment Score and Stock Position.

D. Step-by-step Description:

- Step 1: Taking the input from the user in order to traverse or fetch the data in run time by using the APIs.
- Step 2: The fetched data will be stored in a .csv file and then will be traversed or sort using SQLite.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

- Step 3: Then the regular expression is used for the specificity of data. For the regular expression, the 're' library is used in python.

Eg: import re

If there is a company named 'Microsoft' and you want to check the current position of shares of 'Microsoft' but you don't know the whole name of the company. You just know the company's starting 4 letters. At that time, you can put 'Micr@@@@@' i.e. the letters that you don't know will just be replaced by '@' and the regular expression will provide you all the search starting from the words 'Micr'.

- Step 4: After fetching, the data will be arranged or cleaned properly, i.e. removing the tags, removing the non-letters, and stop words (insignificant words like the, as, to, etc.) using package beautiful soup.

- Step 5: Bag of words will be created after the arrangement of data. Simple numeric representation of a piece of text that is easy to classify is called as 'tokenization'. 'fit.tranform' method is used to fit the model to the bag of words and create the feature vector and the feature vector is stored in an array. This method will require multi-dimensional array. For this, the 'numpy' library is used.

- Step 6: For the data analysis, two libraries will be used i.e. 'panda' and 'numpy'. The comment will be stored in a temporary array where the regular expression will be applied to it from which the words will be picked and bag of words will be formed. There will be some exception for the sarcastic statements that will provide both positive as well negative feedbacks.

- Step 7: The prediction is then made by using Naïve Bayes' Algorithm and the result is written in a new .csv file. There is a library named 'matplotlib' which will create the result for positive or negative feedback. The hash table will be used in order to avoid the error in count if the comment or a review is repeated.

V. CONCLUSION

In this paper, we have shown that a strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company expressed through reviews and news. The main contribution of our work is the development of a sentiment analyzer that can judge the type of sentiment present in the review. The reviews are classified into three categories: positive, negative and neutral. At the beginning, we claimed that positive emotions or sentiment of public in twitter about a company would react in its stock price. Our speculation is well supported by the results achieved and seems to have a promising future in research. Sentimental Analysis provides the ability to analyze the opinions of people for a particular product or for a company. Prediction of stock market is really a hard nut to crack and requires lot of efforts. The market data if analyzed in a proper way can be very effectual in predicting a company's future. We have mined data and trained a neural network to predict the closing price. Though the closing prices are high of a company but due to sentence score, investment in that company will not be a good decision. Instead of investing in a company whose closing prices are high, we will recommend you to invest in a company whose sentimental score is high and positive, there are high chances for its stock prices to go up in future. We have ensured that the error rate while performing all implementation is reducing to the least. This work can be extended for a better output if the data samples are taken for a much longer period.

The method that we adopted to extract the sentiment from the data is by no means robust or dynamic and it most definitely requires more testing and analysis. There are many more areas in reference to this topic where further ideas or methods could be used to contribute to a better output of results. With more time and resources there is a lot of potential for improvements in this area. One of the drawbacks of the analysis was the short range of time being examined, and a larger vole of data is necessary for a more comprehensive return. Another option that the project could expand on could be the use of machine learning for text classification. With this approach the rating system would not be used with the positive and negative word list, but instead the percentage of positive tweets retrieved from the day as a whole would be used as the sentiment rating for each day. A classification algorithm would classify each review as being either positive or negative using a model it created from a set of training data. Once again, with a larger volume of data this could yield good results. However, this depends on how many reviews are made about the stock you wish to analyze.

REFERENCES

[1] Qian, Bo, Rasheed, Khaled, Stock market prediction with multiple classifiers, Applied Intelligence 26 (February (1)) (2007) 2533, http://dx.doi.org/10.1007/s10489-006-0001-7
[2] E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34105
[3] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore
[4] W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011, 5(3): 211-224
[5] J. Bean, R by example: Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation, 2011

[6] Sunil Kumar Khatri, Himanshu Singhal and Prashant Johri.Sentimental analysis to Predict Bombay Stock Exchange Using Artificial Neural Network, Proc. Of ICRITO,2014, pp. 380-384.
[7] Bhat, A.A. and Kamatath,S.S. Automated Stock Price Prediction and Trading Framework for Nifty Intraday Training, 4th International Conference on Computing Communication and Network Technologies (ICCNT),2013, pp. 1-6.
[8] Neethu M.S and Rajasree R. Sentiment Analysis in Twitter Using Machine Learning Technique, 4 ICCC NT,2013,pp. 1-5 Vincent Martin.Predicting the French Stock Market using Social Media Analysis, 8th International Workshop on semantic and social media adaption and Personalization, IEEE,2013, pp. 3-7.
[9] M.P.RajaKumar and Dr. V. Santhi. A Comparision of stock Trend Prediction using Accuracy riven neural network variance.Proc. Of INT, conf. on Control, Commmunication and power engineering
[10] Hui Song,Yingxiang Fan,Xiaoqiang Liu and Dao Tao. Extracting product features from online reviews for sentimental analysis, Computer Science and Converge Information Technology
[11] Xinzhi Wang and Xiangfeng Luo.Sentimental Space Based Analysis of User Personalized Sentiments.Ninth International Conference on Semantics, Knowledge and Grids,2013, pp. 151-156.
[12] Lu Yonghe and Chen Jianhua. Public Opinion Analysis of Microblog Content, Proc.Of International Con.On Information Science and application
[13] Sprenger and Webye.Sentiment Analysis of Stock Market News with Semi-supervised Learning, International Conf. on Computer and Information Science
[14] Bin Wen, Wenhua Dai and Junzhe Zhao.Sentence. Sentimental Classification Based on Semantic Comprehension, International Symposium on ISICS,2012
[15] M.P. Rajakumar, Dr.V.Shanthi. A Comparison of Stock Trend Prediction Using Accuracy Driven Neural network variants,Proc. of Int. Conf. on Control, Communication and Power Engineering, 2013,pp.