Stock Market Prediction Using Sentiment Analysis

Shubham Raj[1], Sindhu Yadav[1], Md. Meraj Alam[1], Vijay Kumar[1], Pruthvi P R[2]

[1]Students, Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.

[2]Assistant Professor, Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India.

Corresponding Author: shubhamrajbest@gmail.com

Abstract: - Social sites like Twitter help millions of people to share their thoughts about the Stock market and what they feel about them. The tweet may be a short and easy sort of expression. Detecting sentiments in-text features a wide selection of applications including identifying anxiety or depression of people and measuring the well-being or mood of a community. Therefore, during this review paper, we focused on Sentiment Analysis of Twitter data. Sentiments are often expressed in some ways, which will be seen like countenance and gestures, speech, and transcription. Sentiment Analysis in text documents is actually a content-based classification problem involving concepts from the domains of tongue processing also as Machine Learning. Using different aspects, the research of Sentiment Analysis of Twitter Data is often performed. We can see the various sorts of Sentiment Analysis and techniques want to perform the extraction of the info. In this paper, we have taken a comparative study of various approaches and techniques of sentiment analysis having Twitter as knowledge.

Key Words: — ***Stock Market Prediction, Machine Learning, Sentiment Analysis, Twitter API.***

I. INTRODUCTION

The most dynamic and advanced means of doing business is the stock market, commonly known as the stock exchange. It is a complex model for little companies, investors and therefore the banking sector to all or any generate revenue and minimize risks. This paper would however plan to use open-source datasets and current data to predict future exchange rates employing a machine-learning algorithm. In the course of years, the share market has been a crucial a part of the expansion of the many companies also as of a country's GDP. In the financial markets of the worldwide private sector, stock markets are given the foremost important position in economic liberalization. There have been a number of flexible impacts on stock markets, the most essential of which are historical data. Many approaches for forecasting stock related data were developed using different techniques and models, which used traditional prices, past revenue growth and dividends, so we all know that we'd like data alongside one among the above factors to effectively predict stocks, in order that the effective market hypothesis are often built.

Manuscript revised July 16, 2021; accepted July 17, 2021.
Date of publication July 18, 2021.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898

In this paper, the Twitter Application Programming Interface (Twitter API), which offers a streaming API, has been taken under consideration within the study of monetary data and continually returns the info. Each data collected reflects the user's status or attitude with reference to a selected subject. This is available between a basic HTTP documentation and a twitter account. After all data is composed for each line, an interpretation is initiated of the emotions relevant to every tweet then a mood is predicted which features a direct impact on the stock status. Sentiment analysis is essentially a drag of classification during which the info content is categorized with a positive or negative opinion. Various models are developed supported various learning algorithms used for the training results. The flowing data is collected through the flowing API after such a model is developed.

II. LITERATURE SURVEY

Twitter may be a popular social networking website where users post and interact with messages referred to as “tweets”. This is a way for people to precise their thoughts or feelings about different subjects. Various different parties like consumers and marketers have done sentiment analysis on such tweets to collect insights into products or to conduct marketing research. With the recent advancements in machine learning algorithms, the accuracy of our sentiment analysis predictions is in a position to enhance. During this report, we had planned

to conduct sentiment analysis on "tweets" using several different machine-learning algorithms. We plan to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked because of the final label.

We use the dataset from Kaggle which was dragged and labelled positive or negative. The info provided comes with emoticons, usernames, and hashtags, which are required to be processed and converted into a typical form. It also must extract useful features from the text like unigrams and bigrams, which may be a sort of representation of the "tweet". We exercise various machine-learning algorithms to conduct sentiment analysis using the extracted features.

III. METHODOLOGY

This project consists of four phases:

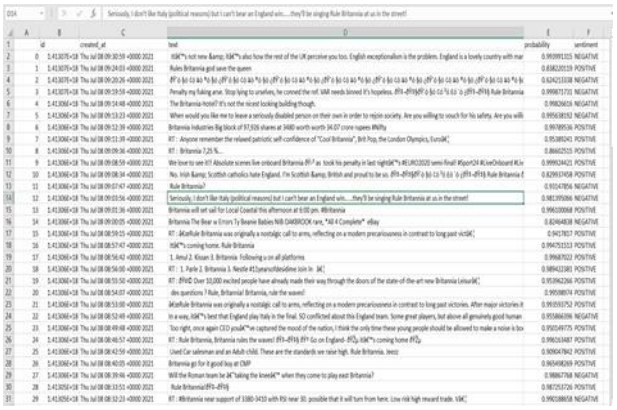
- Data Collection
- Data Processing
- Data Filtering
- Feature Extraction

A. Data Collection

The input data of raw tweets is recovered by using the Scala library called "Twitter4j" which provides a package for real time twitter streaming API. The API requires us to register a developer account with Twitter and fill in parameters like consumer Key, consumer Secret and Token Secret. Twitter API policy allows the developer to extract 50,000 tweets per month. This API extracts random tweets on a particular subject using keywords as filter. Filters supports to retrieve tweets, which match a selected criterion defined by the developer. We used this to retrieve tweets associated with specific keywords, which are taken as input from users.

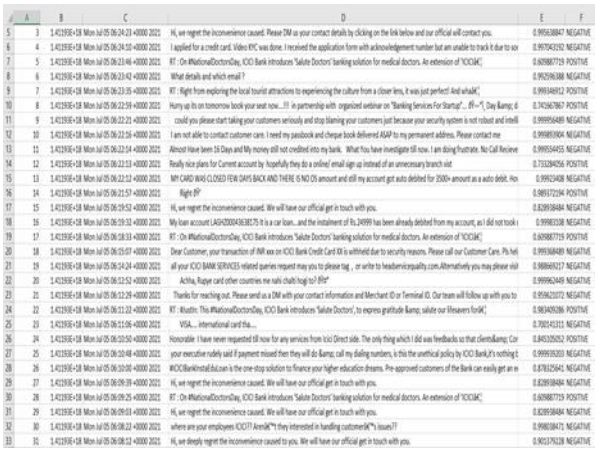
B. Data Processing

Data processing is that the process of splitting the tweets into individual words called tokens. Tokens are commonly divided using whitespace or punctuation characters. It can be single word or cluster of words depending on the classification model used. The bag-of words model is one of the majorly used model for classification. It is supported the very fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest thanks to incorporate this model in our project is by using unigrams as features. It is just a set of individual words in the text to be classified, so, we split each tweet using whitespace. Tweets are simplified by converting it to lowercase which makes its comparison with a dictionary easier.



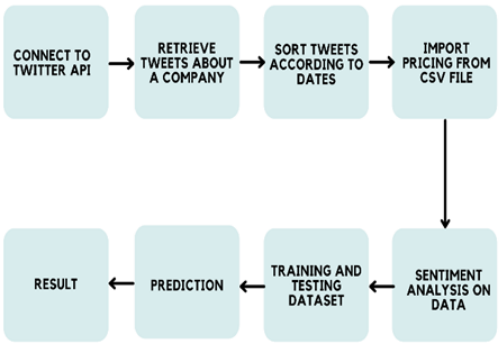
C. Data Filtering

A tweet received after processing still have a portion of raw information in it which we may or might not be useful for our application. Therefore, these tweets are further filtered by pulling out stop words, numbers and punctuations. Stop words: For example, tweets fulfill stop words which are enormously common words like "is", "am", "are" and holds no valuable information. These words set out no purpose and this property is implemented employing a list stored in stopfile.dat. We then compare each word in a tweet with this list and delete the words matching the stop list as Code snippet for stop words removal removing non-alphabetical characters: Symbols like "#, @" and numbers hold no relevance just in case of sentiment analysis and are removed using pattern matching. Regular expressions are used to match alphabetical characters only and rest is ignored. Code snippet for removing non-alphabets this helps to scale back the clutter from the twitter stream. Stemming: it's the method of reducing derived words to their roots.



D. Feature Extraction

This Method Utilized In Text Mining to Seek Out the Importance of a Term to a Document within the Corpus. The Recommended API Is That The Data Frame Based API. This Feature Is Beneficial for a Case Where We'd like to Seek out Trending Topics or to Make Word Clouds. However, Our Project Is More Concentrated Towards Finding Sentiment In Twitter Streams So TF-IDF Is Not Implemented.



IV. Result And Discussion

Twitter API allows 50,000 tweets to be extracted per month for developing purposes.

In our project, in each iteration 100 real time tweets are extracted. The keywords are the specific subjects about what we have to check the prediction.

Keywords like ICICI will work as a filter to extract tweets related to that subject only.

After the filtration and processing module a tweet is becomes a collection of unigrams and multigram.

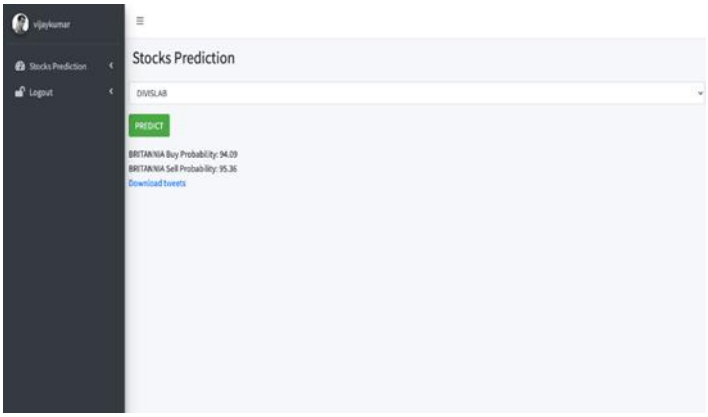
Sentiment analysis id performed on the collected processed data.

After sentiment analysis, each word in the sentence is classified as positive and negative.

After considering the polarity of each word, the polarity of whole sentence is calculated.

The final result is obtained by taking the average of all the negative tweets and all the positive tweets separately.

The probability distribution of polarity about a specific subject gives us the inference whether it is profitable to buy a stock or sell it.



V. Conclusion

During this project, we achieved two goals within the areas of machine learning and interactive visualizations. We developed an honest model of a Twitter sentiment prediction system using supervised text -classification technique and support vector machines. a mix of proven-published ideas (bag-of-words, lexical features, etc.) and novel ideas (polarity buckets, weight-adjusted negation, etc.) contributed to achieving a high-performance system. Several aspects of the system are often improved with a further commitment to the present research. (1) In building a prediction model, a neighbourhood of ensemble machine learning is trending. Similar tasks have reported performance boosts using this system. (2) Hashtags can remarkably help in identifying tweet sentiment. However, since they are compound words, the system was not ready to leverage their full potential in indicating sentiment. (3) We reported a high-performance increase using sentiment lexicons. Since a number of these lexicons are automatically compiled and contain some noise, there is an opportunity for further performance boost if this noise is often reduced. (4) A comparative study of the proposed system on diverse short-text datasets would be required so as to utilize this research in other mediums of sharing sentiments online

References

[1]. Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012).

[2]. Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000-rupee banknotes by Indian government." ICT Express (2017).

[3]. Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.
[4]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015).
[5]. Amolik, Akshay, etal. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016).