A SYNOPSIS ON

Stock Price Prediction Using Sentiment Analysis of News Headlines

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & BUSINESS SYSTEMS

Submitted by:

Shubham Kumar	**A082**
Aashvi Trivedi	**A089**
Jeet Guha Thakurta	**A056**

Under the Guidance of

Guide Name

Designation

Project Team ID: ID No.

School of Technology Management & Engineering

SVKM's NMIMS Deemed to be University

Navi Mumbai

November 2022

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the Synopsis entitled **"Stock Price Prediction Using Sentiment Analysis of News Headlines"** in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Business Systems in the Department of Computer Science and Engineering of the School of Technology Management & Engineering, SVKM's NMIMS Deemed to be University, Navi Mumbai shall be carried out by the undersigned under the supervision of **Guide Name, Designation**, Department of Computer Science and Business Systems, School of Technology Management & Engineering, SVKM's NMIMS Deemed to be University, Navi Mumbai

Shubham Kumar	A082	signature
Name2	A089	signature
Name3	A056	signature

The above mentioned students shall be working under the supervision of the undersigned on the **"Stock Price Prediction Using Sentiment Analysis of News Headlines"**

Signature

Supervisor

Internal Evaluation (By DPRC Committee)

Status of the Synopsis: Accepted / Rejected

Any Comments:

Name of the Committee Members: **Signature with Date**

1.

2.

Table of Contents

Chapter 1

Introduction and Problem Statement

In the following sections, a brief introduction and the problem statement for the work has been included.

1.1 Introduction

The stock market is the backbone of rapidly developing nations. Data mining and artificial intelligence are required for stock market data analysis. Stock price volatility is influenced by the profits or losses of investors in general, or specific businesses. One of the most significant elements that influence the stock market is news stories. A popular objective of investors and organizations is to develop and use a sentiment model to predict by looking at connection between different words using Natural Language Processing by computing results based on previous data collected and labelled, then using the labelled data in Supervised Machine Learning Sentiment analysis model. A key source of information for formulating ideas regarding market investments has long been in the news. Given the abundance of news and its varied sources, it is becoming all the more difficult for an investor or even a group of investors to identify relevant news in the sea of relevant and irrelevant data. Making a sane investment choice at the appropriate time is essential to making the most of the investment strategy. Due to this, computing is essential to automatically receive news from every possible news source, aggregate and filter the relevant ones, and then analyse them to deliver real-time sentiments, both to national and global organizations and businesses that are having a developing or developed status, which are relying heavily on the stock market, only by selling shares to capitalist investors where businesses across the country were able to get substantial capital injections. As a result, the state of our stock market has a direct influence on the progress of a country. Techniques for stock market forecasting are crucial for both inviting new investors to the stock market and retaining existing individuals or organizations. This is due to the general lack of knowledge and awareness, there is a close but implicit connection between the market and the news that is being delivered. Every time new information is made available, it modifies people's perceptions about the organization or the business approach they have adopted. These days, traders and investors constantly have access to the most recent news thanks to the miracles of the internet, which continuously modifies their thoughts and influences their decision to invest in a particular organization The advent of news at any time alters one's

perspective or feeling regarding a specific firm or its established business tactics. Internet acts as a sensation which enables investors and traders now have constant access to the most recent news, which constantly shapes their opinions and determines whether or not they will invest in a particular stock. Computing intervenes to automatically compile news from all pertinent news sources, analyse it, and aggregate it to give users real-time sentiments on whether a stock will increase in value or decrease in value. By using news articles about a firm as the primary source of information and classifying the news as beneficial (positive) and unfavourable (negative), it is possible to forecast the future trend of a stock. The likelihood that the stock price will increase increases when the news sentiment is favourable and decreases when the news sentiment is negative. When news sentiment is positive, the likelihood that the stock price will rise increases, and the likelihood of stock prices will decrease once the news sentiment is negative. The impact of news on stock prices is investigated in this study. We are using supervised machine learning for categorization and other text-mining techniques to assess the polarity of the news. Furthermore, the news must be classifiable and not utilized to build a classifier. There are various classification methods used to evaluate and improve classification accuracy. We used historical data spanning more than five years to carry out this study, thus applying supervised learning models of machine learning. Companies can use these algorithms to adopt reforms that will result in an upward trending stock graph, as a benefit of evaluating news headlines that will help the company perform better and gain a competitive edge, and as a core competency by leveraging sentiment analysis. The data from a variety of sources, including news, social networks, business magazines, and more, can be utilized to acquire consumer feedback on their products and product-related regulations. In this situation, sentiment analysis can be used to procure opinions from people and label them as good, negative, or neutral words, thus it will enable investors to target well-off stocks for the future which will make them invest in their stock early on and gain monetary profit once the sentiment analysis is positive. We will implement the prediction based model using Random Forest count vectors and Naive Bayes probability model.

1.2 Problem Statement

The problem statement for the present work can be stated as follows:

To derive relevant information for investors from a large chunk of Top news headlines using Natural Processing Language to predict the movement of stock market.

Chapter 2

Background/ Literature Survey

In the present times, research work is going on in context of Natural Language Processing and Supervised Machine Learning. In this chapter some of the major existing work in these areas has been reviewed.

Chapter 3

Objectives

In this situation, the objective is to locate the economic news headlines pertaining to the many companies that were previously listed as criteria. This "often abrasive, eye-catching" writing style is used in the title itself to draw readers in and entice them to click through to the entire article.

The project gathers non-quantifiable information about a company from content such as financial news articles and other sources, and then predicts its stock trend based on these headlines using machine learning techniques such as K-means clustering, Decision Trees, Random Forests, Naive Bayes, and Support Vector Machines. All in all, supervised learning techniques are used to forecast stock values based on data produced by sentiment analysis of news headlines. The main objectives of the project can be summed in the following points:-

1. Sentiment Analysis of Stocks -
2. Bidirectional Encoder Representations of Transformers (BERT)
3. Stock Values
4. Predict future stock movement for different stocks

Chapter 4

Possible Approach /Architecture/ Algorithms

We are following a Dictionary based approach which uses Bagging (or Bootstrap Aggregation) technique for text mining to detect sentiment of news articles. This method is based on the research of J. Bean in his implementation of Twitter sentiment analysis for airline companies [1]. Using the collection of both positive and negative words we are going to build the polarity dictionary. Then we can match the article's words against both these word lists and count the number of words appearing in both the dictionaries and calculate the score of that document. We created the polarity words dictionary using general words with positive and negative polarity. For the news article, we are considering the string which contains headline and news body, both. The algorithm to calculate sentiment score of a document is given below.

Algorithm:

1. Tokenize all the headlines in the document into a word vector.
2. Prepare the dictionary which contains words with its polarity (positive or negative).
3. Check against each word whether it matches with one of the words from the positive word from the positive word dictionary or negative word dictionary.
4. Count the number of words belonging to the positive or negative polarity.
5. Calculate Score of the document = count (positive matches) – count (negative matches)
6. If the score is 0 or more, we consider the document as positive or else, negative.

We applied this algorithm using Random Forest using Countvectorizer and Naive Bayes

Possible Algorithms:-

1. Naive Bayes

   It is the one of the simplest and most effective supervised learning algorithms which is used for classification problems. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Bayes' theorem is used to determine the probability of a hypothesis with prior knowledge. Mathematically –

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (4.1)$$

2. Random Forest with Countvectorizer

   During the text classification process, in order to reduce the complexity of text documents and make them easier to work with, the documents have to be transformed from the full text version to a document vector which describes the contents of the document. Countvectorizer is a great tool provided by the Scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for use in further text analysis).

3. Random Forest with TF-IDF

   TFIDF (or tf-idf) stands for ‘term-frequency-Inverse-document-frequency. Where Term-frequency refers to the count of occurrences of a given word in the given document. The problem with using this term-frequency value alone is that-Some irrelevant words (like ‘the‘, ‘and‘, ‘or‘…etc.) occur very frequently in English text documents and these words get higher weightage with respect to term-frequency value while they are not much useful regarding the context of a sentence or paragraph. Inverse document frequency (IDF) on the other hand looks at the presence of a query word among all documents. If a word occurs only in a few documents then it gets a higher IDF value, while if the given word occurs in most of the documents (means not relevant) gets a lower IDF value. In this way, infrequent important words get some highlight and frequent non-useful words are penalised by inverse document frequency

value. Thus it solves the issue with frequently irrelevant words like ‘the‘, but it is still not ideal as it does not rank documents based on the frequency of a given query word. Thus by combining both ‘term-frequency‘ and ‘inverse document frequency‘ statistics it solves both the issues with TF and IDF alone and gives a score value to rank documents based on both. TFIDF score tells the importance of a given word in a given document (when a lot of documents are present). In other words, for a given word query you can actually rank the documents with respect to the relevance with tf-idf score of a term (t), in a given document (d) with respect to a set of documents (D), is defined as-

$$tfidf(t, d, D) = tf(t, d) * idf(t, D) \quad (4.2)$$

Oshi Gupta in her research paper[3] applied all three algorithms mentioned above and the results of –

	Accuracy	Precision	Recall	F1-score
Random Forest with Count Vectorizer	0.851	0.992	0.704	0.823
Random Forest with TF-IDF	0.843	0.899	0.768	0.828
Naïve Bayes	0.851	1	0.698	0.822

4. TextBlob

TextBlob is a python library for Natural Language Processing (NLP) which uses Natural Language ToolKit (NLTK) to achieve its tasks. This tool is used for many NLP related tasks such as language classification, text classification and also for sentiment analysis. By doing a sentiment analysis, we actually determine a polarity value of the sentences, where this value ranges between −1 and 1. Then we label the data with the right sentiment value which is either positive, negative or neutral. For other tools, the polarity value may move on a different scale, so the labelling needs to adjust for these differences for further analysis.

Lazlo and Attila in their research [2] showed that using this algorithm their results were 75.25percent is neutral, 20.25 percent is positive, and 4.50 percent is negative.

5. NLTK -- VADER lexicon

   VADER stands for Valence Aware Dictionary and sEntiment Reasoner. It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

   It is a lexicon and rule-based feeling analysis instrument that is explicitly sensitive to suppositions communicated in web-based media. VADER utilises a mix of lexical highlights (e.g., words) that are, for the most part, marked by their semantic direction as one or the other positive or negative. Thus, VADER not only tells about the Polarity score yet, in addition, it tells us concerning how positive or negative a conclusion is.

   As compared to TextBlob method mentioned in [2], the neutral values have been significantly reduced and we expect that has significant effect in further analysis to obtain more accurate and realistic results with fewer neutral values. 51.50 percent of the total result is neutral in addition to 31.50 percent positive and 17 percent negative. Of the positive or negative categories, the positive strongly dominates, but this huge neutral value still makes the result little bit uncertain.

6. Recurrent Neural Network (RNN)

   Generally when we talk about traditional neural networks, all the outputs and inputs are independent of each other. But in the case of recurrent neural networks, the output from the previous steps is fed into the input of the current state. All in all the Recurrent Neural Network is a neural network that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run. Recurrent neural networks are particularly useful for evaluating sequences, so that the hidden layers can learn from previous runs of the neural network on earlier parts of the sequence.

   One of the advantages using the RNN model is that it can process inputs of any length. An RNN model is modelled to remember each information throughout the time which is very helpful in any time series predictor. Even if the input size is larger, the model

size does not increase. But there is one disadvantage too that due to its recurrent nature, the computation is slow and thus training of RNN models can be difficult.

A significant difference from the previous results of TextBlob[2] and NLTK -- Vader Lexicon[2] is that the neutral section was eliminated, all news headlines were categorized as either positive or negative. The neutral category of the TextBlob was huge, it was significantly reduced by the NLTK -- Vader Lexicon, and then the RNN model was managed to avoid a neutral category. Using the case of AMD, the results came out quite 'balanced' with 51 percent positive and 48 percent negative values.

7. Bidirectional Encoder Representations from transformers (BERT)
   Unlike the traditional NLP models that follow a unidirectional approach, that is, reading the text either from left to right or right to left, BERT reads the entire sequence of words at once. BERT makes use of a Transformer which is essentially a mechanism to build relationships between the words in the dataset. In its simplest form, a BERT consists of two processing models -- an encoder and a decoder. The encoder reads the input text and the decoder produces the predictions. But, because the main goal of BERT is to create pre-trained models, the encoder takes priority over the decoder.

   The results of BERT Algorithm[2] turned out to be 50.50 percent as positive and 49.50 percent as negative.

References

Examples

Journal Paper

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

[2]: László Nemes & Attila Kiss (2021) Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication, 5:3, 375-394, DOI: 10.1080/24751839.2021.1874252

[3] Oshi Gupta, Sentiment Analysis of News Headlines for Stock Trend Prediction, International Journal for Research Trends and Innovation, 2020, Vol. 5, Issue 12, ISSN: 2456-3315

Conference Paper

Text Book / Magazine

Website

[1] Sentiment Analysis of News Headlines for Stock trends (by Oshi Gupta) - https://ijrti.org/papers/IJRTI2012003.pdf

[2] http://www.iraj.in/journal/journal_file/journal_pdf/14-481-153485282984-86.pdf

[3] https://arxiv.org/pdf/1607.01958.pdf