STOCK MARKET ANALYSIS

Article · July 2021

CITATIONS
14

READS
43,234

6 authors, including:

Pallantla Satya Kavya
University of Leicester
72 PUBLICATIONS **323** CITATIONS
SEE PROFILE

Marimuthu Muthuvel
Coimbatore Institute of Technology
29 PUBLICATIONS **201** CITATIONS
SEE PROFILE

Palaniswamy Velvadivu
Coimbatore Institute of Technology
13 PUBLICATIONS **26** CITATIONS
SEE PROFILE

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

STOCK MARKET ANALYSIS USING MAPREDUCE AND PYSPARK

P.Kavya[*1], S.Saagarika[*2], R.T.Subavarsshini[*3], C.G.Nivetheni[*4]
Dr. M. Marimuthu[*5], Dr. P. Velvadivu[*6]
[*1,2,3,4]3rd Year, Msc Data Science (Integrated),Coimbatore Institute Of Technology, Coimbatore, India.
[*5,6]Assistant Professor, Department Of Data Science, Coimbatore Institute Of Technology, Coimbatore, India.

ABSTRACT

Forecasting the stock market has become very important in planning business activities. The prediction of stock price has driven many researches in a variety of disciplines, including computer science, statistics, economics, finance, and operations research. Recent studies have shown that the enormous amount of online information that is available in the public domain, such as Wikipedia, the social forums, news from media, have a significant impact on the investor's opinion towards the financial markets. The reliability of the computational models on prediction of the stock market is very important, because it is highly responsive to the economy and may result in financial losses. In this paper, we have made an extensive analysis on various stocks. First, we have performed Stock Volatility Analysis on 1000 stock dataset of NYSE. The main contributions in this paper include the development of a dictionary-based sentiment analysis model for the financial sector, and the evaluation of the model for scaling the effects of news sentiments on stocks for other markets. By using only the news sentiments, we have achieved a good accuracy of 70.59% in predicting the trends in short-term stock price movement.

Keywords: Dictionary Comparison, Financial Market , News Articles , Sentiment Analysis , Stock Price Prediction.

I. INTRODUCTION

Big data has become a great importance for the expansion of various industries and sectors. It is widely used by business organizations in the formalization of business ideas and intelligence. Furthermore, it has been utilized by the healthcare sector to discover important patterns and knowledge so as to improve the modern healthcare systems. Big data also holds significant importance for the technology, information and cloud computing sector. The financial and banking sectors are utilizing big data to track the financial market activity. Network analytics and big data analytics are also used to catch illegal trading in the financial markets. Similarly, financial institutions, companies, traders and big banks are utilizing big data for generating trade analytics which are utilized in high frequency trading. It also helped in the detection of illegal activities such as financial frauds and money laundering.

In this paper, we hope to build a system which analyses various stocks by using various Big Data frameworks. We are going to use Hadoop MapReduce to find out the top 10 stocks with minimum and maximum volatility and PySpark to predict the closing price of those stocks using Machine Learning models and Sentiment Analysis using news sentiments.

II. PRELIMINARY KNOWLEDGE

Some basic knowledge of MapReduce, Hive and PySpark are reviewed in this section.

a) HADOOP MAPREDUCE

Hadoop MapReduce is a software framework for writing applications in an easy manner which can process enormous amounts of data (multi-terabyte data) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable and fault-tolerant manner. The job of MapReduce is to split the inputdata-set into some independent chunks which are processed by the Map tasks in a completely parallel manner. The framework then sorts the outputs from the Maps, which are then given as input to the Reduced tasks. Generally both the input and the output of the jobs are kept in a file-system. The framework takes care of everything from the scheduling tasks, monitoring them and re-executing the failed tasks. Typically the MapReduce framework (compute nodes) and the Hadoop Distributed File System (storage nodes) are running on the same set of nodes.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

This configuration allows the framework to effectively schedule tasks on the nodes where the data is already there, resulting in very high aggregate bandwidth across the cluster. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker for each cluster-node. This master is responsible for scheduling the job's component tasks on the slaves, monitoring them and re-executing the failed tasks. Then the slaves execute the tasks as directed by the master. Minimally, the applications will specify the input/output locations and supply the map and reduce functions via implementations of appropriate interfaces and/or abstract-classes. These, including other job parameters, contain the job configuration. Then the Hadoop job client submits the job (jar/executable etc.) and configuration to the JobTracker which then distributes the software/configuration to the slaves, then scheduling tasks and monitoring them, providing status and diagnostic information to the job-client. Although the entire Hadoop framework is carried out in Java, MapReduce applications need not be written in Java.

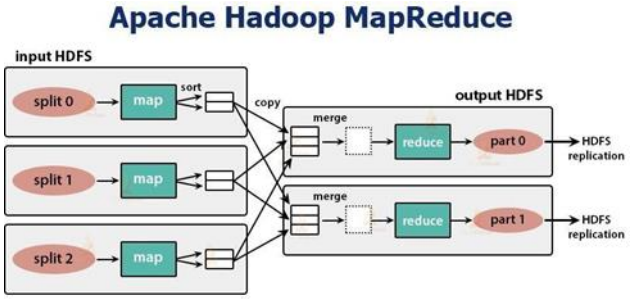
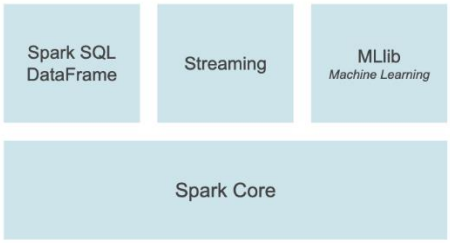


Fig 1: MapReduce Framework

b) PYSPARK

PySpark is nothing but an interface for Apache Spark in the Python language. It not only allows writing Spark applications using Python APIs, but also provides the PySpark shell which provides interactive analysis of your data in a distributed environment. PySpark supports most of Apache Spark's features such as Spark SQL, Mlib (Machine Learning), DataFrame, Streaming and Spark Core.



Spark is a leading tool in the Hadoop Ecosystem. MapReduce with Hadoop can only be used for batch processing and cannot work on real-time data. Spark can work stand-alone or over the Hadoop framework to leverage big data and perform real-time data analytics in a distributed computing environment. It can support all sorts of complex analysis including Machine Learning, Business Intelligence, Streaming and Batch processing. Spark is 100 times faster than the Hadoop MapReduce framework for large scale data processing as it performs in-memory computations thus providing increased speed over MapReduce.

The big-data era has not only forced us to think of fast capable data-storage and processing frameworks but also platforms for implementing machine learning (ML) algorithms that have applications in many domains. With a lot of ML tools available, deciding the tool that can perform analysis and implement ML algorithms efficiently has been a daunting task. Fortunately, Spark provides a flexible platform for implementing a number of Machine Learning tasks, including classification, regression, optimization, clustering, dimensionality reduction etc.

III. STOCK MARKET

a) STOCK MARKET INTRODUCTION

The stock market is nothing but a collection of markets and exchanges where activities such as buying, selling, and issuance of shares of publicly-held companies occur. These financial activities are conducted through

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

regularized formal exchanges or over-the-counter (OTC) marketplaces which operate under a defined set of rules and regulations. There can be any number of stock trading venues in a country or a region which allow transactions in stocks and other forms of securities. If he/she says that they trade in the stock market, it means that they buy and sell shares/equities on one (or more) of the stock exchange(s) that are part of the overall stock market. The main stock exchanges in the U.S. are the New York Stock Exchange (NYSE), Chicago Board OptionsExchange (CBOE) and the Nasdaq. These leading national exchanges, along with several other exchanges operating in the country, form the U.S. stock market. Though it is called a stock market or equity market and is mainly known for trading stocks/equities, other financial securities such as exchange traded funds (ETF), corporate bonds and derivatives based on stocks, currencies, commodities, and bonds are also traded in the stock markets.

b) WORKING

In short, the equity markets provide a safe and controlled environment in which market participants can trade with more confidence in the shares and other eligible financial instruments, with a zero-or low-risk to act in accordance with the rules made by the regulator of the exchange act, as well as primary markets and secondary markets.

The primary market of the stock exchange provides companies with the ability to produce and sell their shares to the public for the first time, at the initial public offering (IPO) process. This activity will help the companies to raise the necessary capital from investors. In fact, it means that the company will be broken up into a series of actions (let's say 20 million shares, and the sale of a portion of the shares (say, 5 million shares to the public at a price (say, $ 10 per share).

In order to facilitate this process, the company will have a market on which the shares are to be sold. This is a market that is defined by the securities market environment. If all goes according to plan, the company to be successful is to sell 5 million shares at a price of $ 10 per share, to raise money for the amount of $ 50 million. Investors will receive shares in the company, they may be expected to hold it for the required period, and in anticipation of growth in the price of the stock, and any potential return on investment in the form of dividend payments. The stock exchange acts as an intermediary in the capital of the process and the compensation for the services of the company and its financial partners.

c) FUNCTION AND PURPOSE

The stock market is one of the best ways to make money for the business, along with the debt markets, which are generally more imposing but not publicly traded. It allows companies to become publicly traded, and the raising of additional capital for expansion by selling shares of the company on the open market. The liquidity of the stock market offers investors a chance to allow their holders to sell securities, fast, and easy. This is an attractive feature of investing in stocks, compared to other less liquid investments such as real estate and other real estate.

History has shown that the value of stocks and other assets is an important part of the dynamics of economic activity, and can influence or be an indicator of public sentiment. In an economy in which the market has been booming, is considered to be an emerging market. The stock market is often a leading indicator of a country's economic strength and development.

d) TAXATION

Taxation is a list of all of the strategies and the profits of the shares, including dividends, and are subject to different tax rates, depending on the type of the safety, security, and longevity. For the most part, the proceeds of the investment in the shares will be subject to capital gains tax. In many countries, corporations pay a tax to the government, and the shareholders pay income tax when they make a profit from owning the shares, which is also known as "double taxation."

e) BEHAVIOR

Changes in the stock prices are mainly caused by external factors, such as socio-economic conditions, inflation rates, and exchange rates. Intellectual capital has no effect on the current gain of the shares of the company. Intellectual capital contributes to the growth of the stock return.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

The Efficient Market Hypothesis (EMH) is a hypothesis in economics, and finance, which states that asset prices reflect all the information that is currently available. The strong efficient market hypothesis does not explain events such as the collapse in 1987, when the Dow Jones Industrial Average fell to 22.6 percent —this is the biggest one-day fall in the United States of America.

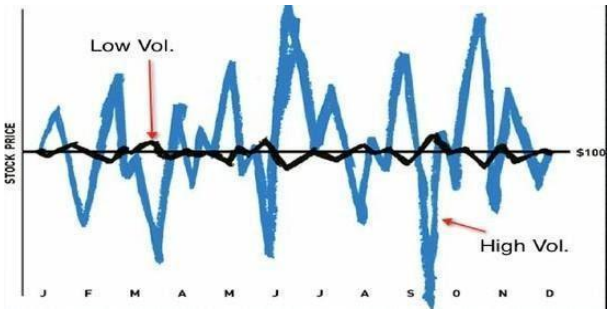
This event demonstrated that share prices can go down, even if there is no one clear reason for this. An in-depth search did not match any 'reasonable' development of events that could explain the attack. (Note that such events are predicted to be strictly random, although very rarely.) In general, it is also true that many of the prices (other than those projected to occur "at random"), are not caused by new information. A study about the fifty largest days to day fluctuations in the U.S. stock prices in the post-war period seems to confirm this.

f) STOCK VOLATILITY

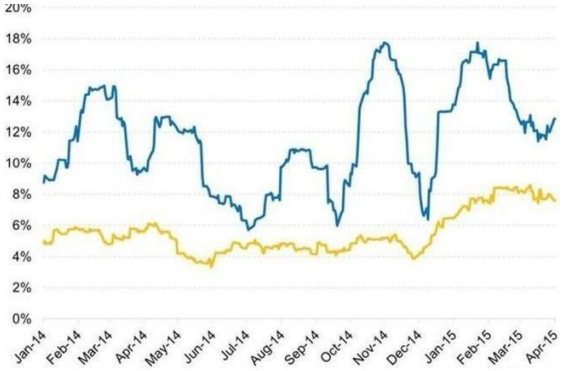
Volatility is defined as the extent to which the price of a security increases or decreases as a result of the win. It is the risk that is associated with a change in the price of a security and is measured by calculating the standard deviation of the annual rate of return over a specified period of time. Simply stated, it's a measure of how quickly the value of an asset or a market is moving.

The variable is usually measured by the standard deviation, or variance. At any rate, the higher the value, the more volatile are the prices or the returns. This means that it has a high standard deviation it means that the prices will be spread out over a wide range of products. Conversely, a low standard deviation indicates that the prices are tightly bound in a narrow range.

- Volatility = swings (varies in a short period of time).
- It is a measure of relative performance of a stock.
- High volatility stocks = more swings and vice versa also.



- Whether the values plummet or rise, the fluctuations depend on a day-to-day basis.



IV. OBJECTIVE OF THE STUDY

- ❖ Stock Volatility Analysis Using Hadoop MapReduce.
- ❖ Stock Sentiment Analysis.
- ❖ Stock Price Prediction.

V. STOCK VOLATILITY ANALYSIS USING HADOOP MAPREDUCE

a) DATASET

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

To calculate the value of the stocks, the volatility we have 1000 sets of the New York Stock Exchange stock information. 1000 CSV files, a CSV file that corresponds to each range. Each of the CSV consists of 7 columns for the Date, Opening, Maximum, Minimum, Close, Volume, and Ad Connect.

b) METHODOLOGY AND IMPLEMENTATION

The standard deviation of the stock volatility, it is now widely used by businesses to identify stocks with high return potential. This is a project that will help you get to the top-10 stocks, with a maximum and minimum volatility. There are three phases of MapReduce jobs:

- **Mapper -1** - The first MapReduce job to read in a CSV file, containing the number of mappers on an equal number of your files. Each mapper will read in the data one line at a time. Each row will contain 7 of the data, i.e., the Date, Open, High, Low, Close, Volume, and the V Shut. However, the computation of volatility only Adj Close and date is required, so first mapper will map set filename (Stock Name) when you turn the key, and a comma-separated-value), date, and adj close value.
- **Reducer -1 -** In this way, the mapping process is complete, the reducer will receive the file name as the key and all values associated with the same key as the iterative values. The Reducer-1, will be, in principle, to compute the Rate of Return for a stock for each month and produce output with key as stock name and single value containing the Rate of Return of a month.

It is used to set the value of the start date and the end date, it is composed of the loading of the initial values of the iterable, and the installation starts from the beginning of the month. Then run a While Loop over the others and take care of the past and current values of each and every moment of each and every month. In fact, the last value. Return will be calculated each time the month changes. The reducer will produce one key value pair for every month of a stock.

- **Mapper -2 -** Next up is the Mapper 2. No significant changes in the data. It will be the core values of the final output of the Reducer -1, and the same data will be sent to Reducer -2.
- **Reducer -2 -** It will take the key as the stock name and then iterable values will have the Rate of Return for all the months for a stock. It will calculate the Volatility of the stock according to the formula, Volatility = Math.sqrt(calc/(noOfMonths - 1) (in the code) and produce the output with key as stock name and values as volatility of the stock.
- **Mapper -3 -** It can be used to sort the data in the figures. It accepts the input data with the same key – value pair as a reducer-2), and the changes made to the key, a value, and the value for a key in order to sort the data by the variable, and pass it according to volatility.
- **Reducer –3 -** The generation of the final solution. This will have input with volatility as key and stock name as value. This will be in sorted order of volatility as the mapper sends data sorted according to the key to the reducer. So now the reducer will just take the first 10 and last 10 stocks which will give 10 stocks with lowest volatility and 10 stocks with highest volatility, respectively, as output. Cleanup method is used to print the top 10 and bottom 10 stocks as it can be done only after completion of the task.

VI. RESULT AND DISCUSSION

Figure 2: These are the top 10 stocks with maximum and minimum volatility.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

VII. MODEL ARCHITECTURE FLOWCHART

Figure 3: Model Architecture Flow Chart

VIII. STOCK SENTIMENT ANALYSIS

In the financial markets, producing a huge amount of data every second." Financial forecasting problems, such as stock tickers, portfolio, education, and risk management are often complex details of the interactions, which are difficult to interpret, or to suggest that, on a purely economic model. The application of deep learning models for this problem can provide more reliable and useful results than the traditional methods. In particular, deep learning is to explore and exploit the patterns in the data that is visible in the present financial and economic theory. Deep learning has a number of advantages in the prediction and classification of permutations that can be avoided, you can take into account the non-linearities and complex data models and the input data, and can be expanded to include all possibly relevant features. In [HPW16], Heaton, et al. presented a variety of deep learning methods with applications in finance, including the show, the model of selection, autoencoders, and lstm center and warehouse. One of the deep learning applications, that is, to see what the use of an autoencoder is to replicate, the replication of a market with a subset of the stores. The consistent nature of these models is a good fit for RNN models such as LSTMs and GRUs. We strive to provide an extension to the existing research in the field of deep time-series analysis as a financial forecast business strategy.

a) DATA COLLECTION

We consider two aspects of the ability to predict the movements of the stock market. The first of these is the social data, reports, and is the second-largest source of information on the historical value of the shares. The News will be downloaded to the New York Times, as well as historical data has been taken from yahoo finance for the stock. In the second place, the daily stock exchange, is information that can be collected on a variety of websites, such as finance.yahoo.com and so on.

b) NEWS COLLECTION

We have collected data from Apple computer, Inc. for the past three years, with effect from January 1, 2013, to April 2, 2016. This data includes the most important events, items, companies, as well as daily stock prices for AAPL in the same time period. Daily stock prices are inclusive of six parameters: the Open, High, Low, Close, Adjusted Close, and Volume. To ensure the integrity of the project as a whole, we have taken into account the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

Individual's closing prices for every day of the stock quote. We've collected data from the largest news aggregators such as news.google.com, reauters.com finance.yahoo.com.

c) PRE-PROCESSING

Text data is unstructured data. Therefore, we cannot guarantee that the raw test data for the classifier is the input data. First, you need to tokenize the document, in words, in order to work at the word-level. Text data that contains noise words, are not contributing to the classification. In addition, the News can contain data, numbers, and more space, plans, punctuation, stop words, etc., etc.

We have to clear the data by removing all of those words. In order to do this, we have created our own list of stop words, which includes specific stop words related to the world of finance, as well as a common English word. This stop words list contains common words, such as proper names, dates, and geographic currency of the songs. In addition, to ignore words which are only one or two of the documents, we will look at the minimum possible number of documents that have to take into account the words that appear in at least three of the documents. The result is the reduction of a word to redundancy. The stemming process, all the words that can be replaced by the original version of the word. For example, with the words 'developed', 'development', 'developing' has been reduced to the origin of the word 'develop'. Some of the preprocessing is done before applying a polarity detection algorithm and some of them are used in the application of the polarity detection algorithms.

d) DEEP LEARNING MODEL

In this study, we use an RNN approach, consisting of lstm (deep learning) to center blocks, to explore the predictive validity of the historical messages of the future of information perception. According to the RNN method, we consider a document as a series of words. If we are to notate in the same order as in the beginning of the text, then this word is denoted by the vector w-t, as described in Section 3.1Here, t runs from 0 to a max sequence length that we define. We introduce a vector of a hidden state as ht , and define a recursion formula as follows:

$$ht = f (Whht−1 + Wxxt)$$

where the function f (x) is the activation function (usually sigmoid function or tanh function), and Wh and Wx are weight matrices. Repeat with the built-in equalizer. Eq.(2) the maximum sequence length, and, finally, the application of a soft-max function, which is to be assigned any value between 0 and 1, giving us the probability of a positive and a negative emotion. The repetition of this process, we update the weight matrices, Wh and Wx. However, the RNN has the problem of vanishing or diverging in the process. The lstm center to solve this problem, the introduction of cells, which decide whether or not to remove or store information. There are four main types of gated cells, that is, an Input gate, a neuron with a self-recurrent connection, a forgot gate and an output gate.

As explained in Section 3.3, we consider the news to be positive, if ri (t) > 0, and negative if ri (t) < 0.) For the purposes of this definition, a stock related to the message length is greater than or equal to 50 is included in the DJ29 in 2013 we get a 16,856 for enjoyment and 17,213 all the negative news.

e) MODEL EVALUATION

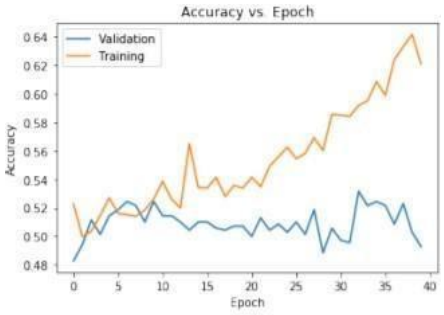
The values of the hyperparameters are set as follows: the batch size is 24, the number of LSTM units is 64, the number of output classes is 2 (positive and negative), the max sequence length is 550 (average length of news), and the number of training iterations is 400 k. Here, we adopted numbers, which are frequently used in the study of movie reviews, for the batch size and the number of LSTM units.

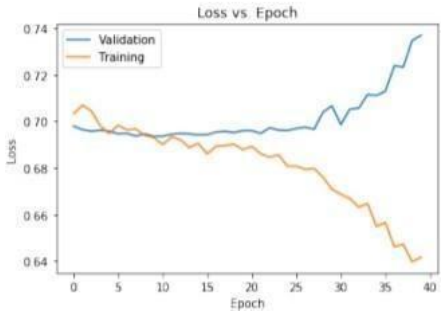
Keeping the threshold as 0.5, the positive and negative sentiments of news are calculated and for AAPL news, we got more sentiments which means APPL stocks might have price hikes.

We show learning curves (loss and accuracy) for our LSTM binary classifier on both the training and validation sets below:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

This shows that the LSTM is very prone to overfitting similar to the convolutional network - as epoch increases, so does training accuracy, but validation accuracy stays constant, or validation loss increases.





IX. STOCK MARKET PREDICTION

One of the ways to predict stock is to use the concept of using machine learning algorithms. We use it with the help of Spark which is one of the major tools of the Hadoop Ecosystem. Spark is an open source, library program used for big data projects.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

a) DATA COLLECTION

One of the ways to predict stock is to use the concept of using machine learning algorithms. We use it with the help of Spark which is one of the major tools of Hadoop. Ecosystem. Spark is an open source, library program used for big data projects. Data collection is one of the most important steps to accurately predict stock market behavior. There are two ways to collect one data by analyzing daily stock market data and the next is to find calendar information. Everyday details in the stock market can be collected on many websites like finance.yahoo.com etc. Also, stock market information can be downloaded through many APIs.

b) PRE- PROCESSING NUMERIC DATA

We have taken the stock prediction dataset which has an independent variable x i.e., Date, Open, High, Low, Close, Volume and one dependent variable y i.e., LABEL**.**

c) FEATURE SELECTION

The ISNULL () operator is used to process the return value if the value in a particular row or column is null. Performs functions such as mean, median, and mode and returns the solution of the sentence if the expression is not null. But in terms of our data, there is no need to replace null values with new values because the percentage of lost value is about 2%. We therefore plan to drop the data. Since dumping this 2% data does not make a change in the model. Therefore 2% of the data is reduced using dropna () function.

```
root
 |-- Date: string (nullable = true)
 |-- LABEL: integer (nullable = true)
 |-- Open: double (nullable = true)
 |-- High: double (nullable = true)
 |-- Low: double (nullable = true)
 |-- Close: double (nullable = true)
 |-- Volume: double (nullable = true)
 |-- InterestRate: double (nullable = true)
 |-- ExchangeRate: double (nullable = true)
 |-- VIX: double (nullable = true)
 |-- Gold: double (nullable = true)
 |-- Oil: double (nullable = true)
 |-- TEDSpread: double (nullable = true)
 |-- EFFR: double (nullable = true)
```

Our next step is to choose the attributes that are necessary and then leave the attributes that we don't need. We have selected variables based on the output of the correlation matrix. The scale is best used for variables that show a linear relationship between each other. It can be indicated by a scatter plot. With the help of this, we have found the required variables. The selected variables are "LABEL", "Open", "High", "Low", "Close", "Volume", "Interest rate", "ExchangeRate", "VIX", "Gold", "Oil", "TEDSpread". We selected it by using data.loc [] function.

```
+-----+-----------+-----------+-----------+-----------+--------+------------+------------+----+----+-----+---------+
|LABEL|       Open|       High|        Low|      Close|  Volume|InterestRate|ExchangeRate| VIX|Gold|  Oil|TEDSpread|
+-----+-----------+-----------+-----------+-----------+--------+------------+------------+----+----+-----+---------+
|    0|12266.63965|12659.82031|12266.46973|12654.36035|2.9553E8|        1.77|      1.5615|22.68|897.0|100.92|      1.3|
|    1|12651.66992|12696.29004|12555.16992|12608.91992|2.3276E8|        1.72|      1.5618|23.43|893.5|104.83|     1.31|
+-----+-----------+-----------+-----------+-----------+--------+------------+------------+----+----+-----+---------+
only showing top 2 rows
```

The next process is to split the training and testing data. We are implementing this by using the randomSplit() function . In this we are splitting the data into the ratio of 7:3.

To improve the preprocessing process, we divide the attributes into 2 sets such as numCols containing numerical values, and catCols which have categorical variables. We use One hot encoding process in catCols, which helps us better represent category information. This is because most machine learning algorithms do not work better with categorical variables, so it is better to encode. One hot code entry will be 0's and 1's.It will make them a standard state to identify whenever it is required or needed while performing.

Label encoding is used to convert categorical values into numbers so that the machine is able to read the data. It is considered an important step in the preprocessing of structured data. Each machine learning algorithm determines how well the labels work. In labeling, we will place the variables in categories between 0 values and (number of classes - 1). Of the selected values, all of them were converted into numerical values except height which has already consisted of numerical values by using the function **LabelEncoder ().**

String Indexer helps to provide label indices to machine learning columns. It is nothing but when we give the input column in numbers it changes that into strings and then it indexes the string values.

To feed the columns into the machine learning algorithms the entire column list needs to be converted to a single vector value. It is considered helpful in combining various features into a vector of a single element. This is mainly done to train ML models such as SVM, random forest, Decision Tree etc. It is done using the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

VectorAssembler () function. We are creating a new column named VectorAssembler_feature and inserting the output into this column.

All the input features provided are converted into a single object using a function called Dense Vector function. The data framework has 11 columns and features that are now ready are given as input into the machine learning algorithm.

h) MACHINE LEARNING MODEL:

Machine learning algorithms are performed by Spark which has a function named Mlib. Then we fitted the Decision tree regressor to the training data set. Here the aim is to find the label values to determine whether the stock is increasing or decreasing at the end of a particular day. We have chosen this because the results are better on this when compared with other models and it also reduces the overfitting of training variables. We have got 90% accuracy by implementing the Decision tree algorithm.

g) MODEL EVALUATION:

In this case, MLs and big data tools have been used for better stock market analysis. Because the stock market is often done. For evaluating the model, we used 2 metrics, first is a Mean absolute (MAE) error. It is a measure of the difference between paired observations which is expressed in the same way. It helps to measure the accuracy of the continuous variance. Our predicted MAE score is 1.024, which is a good score.

The second metric counts the r-square ratio, the ratio that shows the value of the x variance (dependent variable) which is defined by the independent variable. A value of more than 60% r square is considered positive. In our model, the score rate is 70%.

X. CORRELATION ANALYSIS OF PRICE AND SENTIMENTS

After training and testing the model, we get coefficients which will tell us basically how the stock market price is related to the headlines and the historic price of the particular stock. The positive coefficients determine that the two factors are directly related and negative coefficients tell us that data is inversely proportional. This correlation model is going to help us to predict the stock price of the new data set.

XI. LIMITATIONS OF THE STUDY

- One of the major limitations of this approach is the changing stock market prices. Although the model predicts the stock of the next prices, there are additional factors from nature that affect stock prices.
- Using technical analysis to select stocks and increasing your investing potential to invest in which stock through algorithms, may lead to risk management and might tend to lose your money.

XII. FUTURE SCOPE OF STUDY

- Potential improvements can be made to the data collection and analysis method.
- Future research can be done on potential improvements such as, using more refined data, time frames and more accurate algorithms that are associated with the new dataset.
- Real time trading model, with live streaming data that can be upgraded to calculate the total returns or investments in real time. This helps in increasing the efficiency of the model and would boost the accuracy of the model.

XIII. CONCLUSION

- In Stock Volatility Analysis, being below the bottom of the numerical range, zero is obviously a very special form of volatility. Volatility is zero if there are no changes in the price (price remains constant). So, the above mentioned are top 10 minimum stocks whose price would never change, since it is closer to zero.
- Higher volatility means that a collateral value can potentially be spread out over a larger range of values. So, the top 10 are those maximum stocks whose price of the security can change significantly over a short time period in either way. A lower volatility means that a collateral value does not fluctuate dramatically, and tends to be steadier.
- In this case, ML and big data tools have been used for better stock market analysis. Because the stock market is often uncertain. With this, we are able to avoid the investors from facing significant financial losses.

- From this, it is clear that this process of forecasting the stock market has been quite difficult and many preprocess steps have been done to bring high accuracy to the model.
- From the Machine learning predictions, we are able to predict the stock market movements which helps the investors to invest money at the correct place and in a timely manner. The calculated results show that the Decision tree works better than the SVM and logistic regression. Apart from these regression methods, many methods can be implemented in finding better results. This can be achieved by using the Neural networks, by increasing the number of nodes for better accuracy. And we also found that the Feed-Forward neural network gives a more accurate prediction of the opening price of the stock.
- This study is also a niche application of sentiment analysis in gauging the effects of news sentiments on stocks for the pharmaceutical market sector. One major contribution of this work is a sentiment analysis dictionary.
- The sentiment scores obtained from the analysis of the news articles is a powerful indicator of stock movements and can be used to effectively leverage the prediction of short-term trends. We believe that the reason the model is able to achieve an accuracy of 70.59% with the dictionary-based approach is that the dictionary was specifically created for this particular sector by researching and leveraging domain expertise.
- By analyzing the results of sentimental analysis and prediction, the stock of Apple is positive and good. We therefore recommend investors to invest in Apple and make a good profit.

XIV. REFERENCES

[1] Sentence Embedding Prediction.Ipynb - Used The Event Embeddings Into An Lstm.

[2] Https://Github.Com/Srizzle/ Deep-Time-Series/Papers - Relevant Papers (Used For References)

[3] Https://En.Wikipedia.Org/Wiki/Stock_Market

[4] Https://Www.Investopedia.Com/Terms/S/Stockmarket.Asp

[5] Https://Www.Analyticsvidhya.Com/Blog/2021/01/Image-Classification-Using-Convolutional-Neural-Networks-A-Step-By-Step-Guide/

[6] Https://Medium.Com/@Raghavprabhu/Understanding-Of-Convolutional-Neural-Network-Cnn-Deep-Learning-99760835f148

[7] Http://Www.Iraj.In/Journal/Journal_File/Journal_Pdf/14-481-153485282984-86.Pdf

[8] Https://Arxiv.Org/Ftp/Arxiv/Papers/1607/1607.01958.Pdf

[9] Https://Github.Com/Adityajain10/Stock-Predictor-Pyspark-Mlib

[10] Https://Www.Analyticsvidhya.Com/Blog/2019/12/Streaming-Data-Pyspark-Machine-Learning-Model/

[11] Multistep Pred.Ipynb - Used Vader Sentiment And Previous Prices For An Lstm