International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 25th Jun 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 05

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 05/25

Title **Stock Movement Forecasting Using News**

Volume 11, SPL ISSUE 05, Pages: 163-169

Paper Authors
Mr. Aggala Chiranjeevi, Praveen Vipparthi , Venkata Pavani, Aparna Pudi , Sri Sai Sravan

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Stock Movement Forecasting Using News

Mr. Aggala Chiranjeevi[1], Praveen Vipparthi[2], Venkata Pavani[3], Aparna Pudi[4], Sri Sai Sravan[5]
[1]Asst. Professor, Dept of CSE, [2]18ME1A0582, [3]19ME5A0504, [4]18ME1A0 584, [5]18ME1A0572
Ramachandra College of Engineering, A.P., India
aggala.chiranjeevi@gmail.com , praveenvipparthi007@gmail.com, pavanigullapudi030@gmail.com, aparnapudi33@gmail.com, sravanmusunuri@gmail.com

ABSTRACT
Without a question, stock markets are an important and necessary aspect of every country's economy. However, the influence of stock markets on a country's economy may differ from the impact of stock markets on other nations' economies. For many investors, the stock markethas become an appealing investmentdestination, and it has grown tremendously over time. However, due to the stock market's volatility, many investors are hesitant to invest in it. Investors in a huge stock market frequently take risks, and they are concerned about losing their hard-earned money. However, because the stakes in the stock market are so high, investors must be willing to take the same level of risk, and they must beconfident in the investing approachthey chose to assure maximum return. This study uses the fundamental analysis approach to determine a stock's future trend by analysing news items about the firm as primary data and attempting to classify the news as good (positive) or poor (negative). If the news sentiment is favourable, thestock price is more likely to rise and ifthe news sentiment is negative, the stock price is more likely to fall. In this project, we use Time Series Analysis for predicting the movement or trend of future stock price i.e if it goes up or down the next day. We create a dataset containing stock data and News_Factor, which is created for representing the polarity of news of that day.

Introduction

Stock movement is mostly based onthe public opinion of stocks. If people believe that the stock price will increase in the future, then they will invest in that company by purchasing shares of that company. So, at any point of time, if the stock price of a company increased, it means that the number of buyers > sellers and vice versa is true. We need a major factor which represents the public opinion and could also change it. The answer is News. The summary of a complete article can be found in the title itself. So, instead of using the whole article we use titles of news articles to our advantage in this project.

We use news articles from CNBC.com I ECONOMICTIMES.com, etc which are very popular for their credibility. But since they do not maintain news archives sorted by date, we make use of Google News.

We specify the website we want the results from and also the time range in the google news section. We use Web Scraping to extract the date and title of all articles in the page. We use Automation to repeat this process to all the pages by automatically simulating a mouse click on the next option.

We use pre-trained NLP models to gain

insight from the news title in the form of Polarity. We use this number as the News_Factor mentioned previously.

The dataset could be prepared using the fields: date, stock price, news_factor and the output variable, in this manner. This dataset is fed to the Time Series Analysis model and the output variables are recorded for calculating accuracy.

RELATED WORK

In their study [1,] Nagar and Hahsler proposed an automated text mining-based technique for aggregating news articles from multiple sources and generating a News Corpus. Natural Language Processing (NLP) methods are used to filter the Corpus down to useful sentences. As a measure of the sentiment of the whole news corpus, NewsSentiment, a sentiment metric based on the count of positive and negative polarity terms, is presented. The news collecting and aggregation engine, as well as the sentiment evaluation engine, were built using a variety of open source packages and tools. They also claim that NewsSentiment's temporal fluctuation has a good association with actual stock price movement.

According to the research paper [2], They have made use of KNN (Kth nearest neighbour) machine learning algorithm. KNN is a powerful mathematical model which is very good at detecting mathematical patterns hidden within data.

Problem Statement

A. Existing Model

In the existing model, the prediction is completely based on the numerical patterns within the data. KNN is used to make the prediction. But statistical analysis is insufficient in making the decision. Hence, there is scope for improvement if we include other real world aspects on which stock movement depends.

Disadvantages

Real world events which could alter the stock movement are not considered in the existing model.

Prediction that is made completely based on the statistical and numerical analysis cannot be considered accurate.

B. Proposed Model

In the proposed model, we make use of news which represents events in the form of news articles. We make use of news titles which gives us the gist of the article. Here, a time series analysis model is used to make the prediction. Time Series Analysis is chased because the future data points within the dataset depend on the past ones.

Since the output field within dataset isa binary field, has 2 possible values, it comes under directional forecasting.

Traditional machine learning is better at directional forcasting. That is why a Neural Network based model is not chosen here.

Advantages

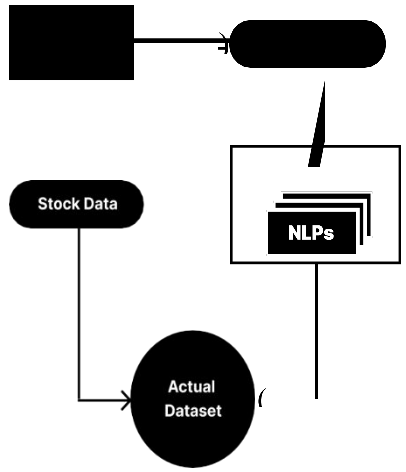
The prediction that is made is based on both numerical and news data analysis. Here, the actual events that altered the stock movement/trend are taken into account.

Large number of titles from various sources can be fetched since automation and web scraping are used.

Implementations Details

Dataset Preparation

The flow of data for dataset preparation is shown in *figure 5.1.*



First, we need to get the titles dataset Using, Automation and Web Scraping Since many news articles are published online, we can access them using Google News.

Automation & Web Scraping

For using Automation Tools like *Selinium,* we need a file called *WebDriver* which allows *Selinium* to control the web browser. Using this, we can open any URL in the browser and wait for the content to load on the browser. Then, after all the dynamic content has been loaded, we could get the loaded HTML of the page.

Using Web Scraping tools like *Beauitifu/Soup,* the retrieved HTML can be parsed and the required information can be extracted by using *class* names *id* names of specific elements that we want to target like date & title.

Titles Dataset

This dataset contains dates and their lists of respective titles. Some dates may not contain any titles at all. Those dates will be filled with *None.*

Sentiment Analysis

Sentiment analysis is a type of text mining that discovers and extracts subjective information from the source material. The classification of the polarity of a given text at the document I sentence, or feature/aspect level, whether the conveyed opinion in a document, a sentence, or an entity feature/aspect is positive, negative, or neutral, is a basic job in sentiment analysis. For example, advanced "beyond polarity" sentiment categorization examines emotional states such as pleasure, rage, disgust, sadness, fear, and surprise.

In this research, we use multiple pre-trained NLP models for gaining insight and finding out the polarity of titles.

Using multiple pre-trained NLPs is required because sometimes polarity of titles may not be recognised by a certain NLP but can be done by another. To cover for this disadvantage, we use multiple NLPs and consider only the maximum polarity value.

In this research, we use two pre-trained NLP models. They are VADER, TextBlob. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is tuned in to social media sentiments.

VADER employs a mix of techniques. An emotion lexicon is a collection of lexical characteristics (such as words) that are classified as positive or negative depending on their semantic orientation. VADER not only displays the Positivity and Negativity scores, but also the degree to which a sentiment isfavourable or negative.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

TextBlob is a Natural Language Processing Python module (NLP). Natural Language ToolKit (NLTK) was used extensively by TextBlob to complete its objectives. NLTK is a library that allows users to deal with categorization, classification, and a variety of other tasks by providing simple access to a large number of lexical resources. TextBlob is a basic package that allows for extensive textual data analysis and processing.

A sentiment is determined by its semantic direction and the intensity of each word in the phrase in lexicon-based techniques. This necessitates the use of a pre-defined vocabulary that categorises negative and positive terms. A text message is often expressed by a collection of words. Following the assignment of individual scores to all of the words, the final emotion is derived using a pooling technique such as the average of all the ratings.

The polarity values of both VADER & TextBlob are between 1 & -1. Positive numbers represent positivity and vice versa. O represents a neutral statement. Subjectivity values can also be considered if required. They refer to the magnitude of personal feelings within the statement passed. But in this research, subjectivity is not used. Polarity is considered.

Stock Data

Stock data is fetched from the *yfinance* module in Python. *yfinance* is a well-known open source library created by Ran Aroussi to access the financial dataon Yahoo Finance.

Yahoo Finance has a wealth of market information on stocks, bonds, currencies, and cryptocurrencies. It also provides market news, research, and analysis, as well as options and fundamentals data, which distinguishes it from its competitors.

Stock dataset contains a total of 6 fields: ***Open, High, Low, Close, Adj Close, Volume.***

Open refers to the starting stock price of that day. High refers to the highest point of price during the day. ***Low*** refers to the lowest point of price during the day. ***Close*** refers to the stock price at the end of the trading time of that day.

Volume refers to the number of shares traded during the trading time of that day. ***Adj Close*** or ***Adjusted Closing Price*** is the closing price after accounting for any corporate actions that took place on that day. For example, deducting the dividend amount per share from the closing price of the stock gives that adjusted closing price of that day.

The Stock Dataset that we consider in this research does not contain all the above fields. Since we need a single number which represents the stock price of the whole day during trading, we take the average value of ***Open*** & ***Close.***

$$\text{Value}= \frac{(\boldsymbol{Open} + \boldsymbol{Close})}{2}$$

Actual Dataset

The data fields in the actual dataset that we consider for Time Series Analysis is shown below.

Date	Stock Value	News_factor	OUTPUT

News_factor is the average value of all the highest polarities of all the news titles on that day.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

$$\text{News\_factor} = \frac{\sum_{n}^{i=1} \max_{abs}(V(S_i), T(S_i))}{n}$$

where Si refers to the ith sentence, V( ) refers to the VADER polarity function, T( ) refers to the TextBlob polarity function.

The function *maxabs( )* will convert the values returned by both the functionsto absolute values for getting the maximum. Absolute values are considered only for compansIon purpose. *maXabs()* will return the direct value returned by the functions.

Computing average helps in finding out if the overall news is positive or negative on that day.

The OUTPUT field contains one of the 2 values i.e -1 & +1. -1 refers that the stock movement is downwards. +1 refers that the same is upwards. Since the output is binary, this comes under directional forecasting. Directional forecasting is better achieved using traditional machine learning than deep learning techniques. So, we have used Time Series Analysis model called ARIMA for making the actual prediction.

Time Series Analysis

A time series is just a collection of data items that occurred in a specific order over a period of time. The only assumption is that the process is "stationary," meaning that the beginning of time has no impact on the features of the process under the statistical component. TSA is the foundation for time-based problems and forecasting analysis. Time series analysis is a method for studying a collection of data points over a period of time. Instead of capturing data points sporadically or arbitrarily, time series analyzers capture data points at constant intervals over a predetermined length of time. This form of analysis, however, is more than just gathering data over time.

The ability to depict how variables change over time distinguishes time series data from other types of data. In other words, time is an important variable since it reveals how the data changes through time as well as the ultimate outcomes. It provides an extra source of data as well as a predetermined order of data dependencies.

To maintain consistency and dependability, time series analysis often requires a high number of data points. A large data collection guarantees that your sample size is representative and that your analysis can cut through noisy data. It also guarantees that any found trends or patterns are not outliers and that seasonal variation is taken into consideration .

Box-Jenkins ARIMA models: These univariate models are used to better explain and forecast future data points for a single time-dependent variable, such as temperature across time. The assumption behind these models is that the data is stationary. Analysts must account for and eliminate as many variations and seasonalities as possible in previous data points. The ARIMA model, thankfully, includes terms to account for moving averages, seasonal difference operators, and auto regressive terms.

Box-Jenkins Multivariate Models: Multivariate models are used to examine many time-dependent variables throughout time, such as temperature and humidity.

Holt-Winters Method: The Holt-Winters method is an exponential smoothing technique. It is designed to predict outcomes, provided that the datapoints include seasonality.

The qualities of a stationary time series are independent of the time at which it is viewed. Thus, time series with trends or seasonality are not stationary; the trend and seasonality will impact the time series' value at various periods. A white noise series, on the other hand, is stationary - it should seem the same no matter when you look at it.

Logarithms and other transformations can aid in the stabilisation of a time series' volatility. By removing fluctuations in the level of a time series and so eliminating (or lowering) trend and seasonality, differencing can assist stabilise the mean of a time series.

ARIMA

An **Autoregressive Integrated Moving Average** model is a type of regression analysis that determines how strong one dependent variable is in comparison to other changing variables. The purpose of the model is to anticipate future securities or financial market movements by looking at the discrepancies between values in a series rather than actual values.

A model that displays a changing variable regressing on its own lagged, or prior, values is known as **autoregression** (AR).

Integrated (I): denotes the separating of raw observations in order for the time series to become stationary

Moving average (MA): A moving average model applied to lagged observations combines the dependence between an observation and a residual error.

ARIMA Parameters:

ARIMA treats each component as a parameter with a consistent nomenclature. ARIMA with p, d, and q is a standard notation for ARIMA models, where integer values replace the parameters to denote the kind of ARIMA model utilised. The parameters are as follows:

p - Lag Order

d - Degree of Differencing

q - Order of Moving Average

$\mathrm{ARMA}(p', q)$ is given by:

where,

L - Lap Operator

α_i - parameters of autoregressive part θ_i - parameters of moving average part

The above part can be written according to ARIMA as:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t.$$

The above expression defines ARIMA(p,d,q) along with drift as shown below:

$$\frac{\delta}{1-\sum\varphi_i}$$

Testing

To test Time Series Analysis Models, generally the test dataset is divided into *80:20* ratio for training and testing. After training the model, the test dataset is passed for prediction and the accuracy is calculated based on the percentage of correct predictions out of the complete test data. But this method is not right for Stock Trend Prediction using News

Fig 8.1

Consider the above figure 8.1.
The figure depicts the complete dataset from the year 2010 - 2022. The dataset is divided in the ratio *80:20* as shown above. If the training dataset comprises of data from 2010 - 2019, then the model does not have any information about the News_factor of data present in the test dataset. This defeats the purpose of having News_factor. The prediction from this model will not depend on the recent news of the test data. Hence, it does not produce accurate reading.

To overcome this problem, we need a testing strategy that would take recent news into account.

Consider fig 8.2. In the figure, *N days* represent the complete dataset that we have. (N+1)th day is the only day which we could predict with complete accuracy. So, to calculate accuracy of our model, we

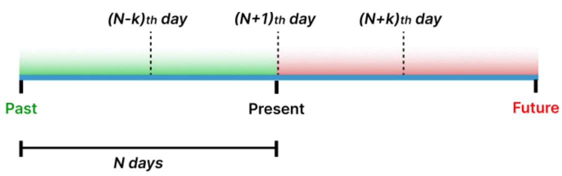


Fig. 8.2

need to use past data within the dataset as well. So, if we want to predict the *Movement* on (N-k)th day, then we have to consider all the data prior to this day i.e until (N-k-1).
The same process can be done for different k values and the accuracy can be calculated by comparing the results of the prediction with actual movements within the dataset.

The disadvantage of this method is that we have to re-train the model for every testcase . But this is the only method in which we could get maximum accuracy.

Conclusion

When the model is tested using the Apple (AAPL) stock data using the above strategy, 85% accuracy is achieved, which is significantly higher than the existing model. A 15% increase in accuracy is observed. Hence, using Time Series Analysis proved to be a very efficient tool at predicting stock movement using news.

References

[1] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from News, IPCSIT vol. XX (2012) IACSIT Press, Singapore

[2] R.S. Latha, G.R. Sreekanth, R.C. Suganthe, M. Geetha, R. Esakki Selvaraj, S. Balaji, K.R. Harini, P. Priya Ponnusamy, Kongu Engineering College, Erode, India

[3] www.analyticsvidhya.com/ blog/2021/10/a-comprehensive- guide-to-time-series-analysis/

[4] https://en.wikipedia .org/wiki/ Time_series

[5] www.machinelearningplus.com/ time-series/arima-model-time-series-forecasting-python/

[6] www.adityabirlacapital.com/abc-of-money/factors-affecting-stock-market