Analyzing Stock Trend using News Articles

Shravan Bhat, Siddhanth M, Sampath Kumar,
Dr. Rekha B Venkatapur, Head of the Department
Dept. of Computer Science & Engineering
K.S Institute of Technology, Bangalore, India

Abstract **- "Data" surpassed "Oil" as the most valuable resource in the world. We are living in an age where the value of "data" is more than any other resource. As such, the world economy is in one way or another linked to the data that is being produced. The world economy runs on the basis of the stock market. The stock market is intertwined with the current affairs and the news. For instance, the news of bad loans in the crisis of Yes Bank, as it dropped by 86%. This is an example of how news affects the stock market. There are many factors by which the stock trends are affected, one of which is daily news articles.**

Recent studies have shown that the massive amount of online information and various social media discussions and news stories tend to have an observable effect on the financial market. So, the goal would be to analyze and determine whether there is any significant link between the news articles and the news on the internet on the stock market or rather whether it has any impact on the shares of stocks of a company

Keywords - Machine Learning, Natural Language Processing, Stock market prediction, Analytics, Neural Network

I. INTRODUCTION

Stock market is an aggregation or a cluster of buyer and seller of stocks, which basically represent the ownership of a business. So, these stocks can be bought and sold on stock exchanges. Since, the stocks issued by individual companies are affected by many different factors both inside and outside the company, the stock market is very unpredictable. Therefore, a successful prediction could yield a significant profit. Recent studies have shown that the massive amount of online information and various social media discussions and news stories tend to have an observable effect on the financial market. So, the goal would be to analyze and determine whether there is any significant link between the news articles and the news on the internet on the stock market or rather whether it has any impact on the shares of stocks of a company. We can also thus figure out how each news headline could in turn change the stock market.

II. METHODOLOGY

The project is broken into 6 parts
PART I: Data collection and sentiment analysis
PART II: Developing the ML model
PART III: Training the ML model with training data
PART IV: Calculating the performance of the model.
PART V: Testing the ML model with testing data
PART VI: Accuracy of the ML model.

A. Data collection and sentiment analysis

First step is to download the data from various news sources and their respective api's. The news sources we used for retrieving the data are:

1. https://www.economictimes.com
2. https://www.deccanherald.com
3. https://www.moneycontrol.com
4. https://www/finance.yahoo.com
5. https://www.investing.com

We processed over 20 lakh news articles over 8 years which is more than any previous study that we could find. Data is downloaded from the stock market indices and platforms with information like high, low, volume traded etc. This scraping of information will be done with help of BeautifulSoup4 - A library in python for extracting data. We will now parse the given information which has been downloaded, to process and remove any unnecessary information. From the news articles, only the financial news will be loaded and any extra tags or information will be discarded. The relevant fields from the stock market data will also be parsed in similar manner.

<u>Determining the polarity of the news article</u>
This is done by using the library Vader Analysis. The library goes through the article and assigns a value which is used in determining the polarity of the news article. Vader library is used for determining polarity in a very efficient way. The library classifies information into 4 different types:

1. Positive: if the assigned score > 0
2. Negative: if the assigned score < 0
3. Neutral: if the assigned score ~ 0
4. Compound: the sum of positive and negative and the sentiment score

After downloading the news articles, we assign each heading a vader score. We chose this library since we found it has a lot of accuracy for news articles. The negative aspect is that for financial news articles there tended to be more false positives due to which we also had to include some bag of words for common negative sentiment words which were being wrongly classified. Thus we achieved a parsed csv file which had the vader score for each headline.

B. Developing the ML model

The ML model was developed using Tensorflow and Keras library and was executed on Google Colab. The dataset is divided into 80-20 ratio (80 for training, 20 for testing)

The ML model can be broken down into 6 parts:

a) Importing all the dependencies
b) Creating the neural networks
c) Training the model
d) Evaluation of the model
e) Testing the model
f) Accuracy of the models.

C. Developing the ML model

After developing the ML model, it is compiled and then trained. The training process involves using the tensorflow library with keras. The code for running the Ml model is as follows:

```
train_model=model.fit(X_train[0:],y_train[0:],
epochs=500,verbose=False,shuffle=True)
```

where epochs represents the number of iterations or the total number of samples on which the model is training on. Here, the number of samples used are, 79,000*500 = 39500000 samples
fit() is the method used in Tensorflow to invoke the training process

D. Training the ML model with training data

After training the machine learning model i.e after processing samples, the results of the machine learning model are then analyzed for performance. The model is then taken to the next stage i.e testing the model with the testing data.
A snapshot of the percentage change of the shares of stock according to the model:

```
Input: [[-0.232 -0.2 -0.23 ]]
Prediction: -0.05841918
```

In the snapshot, the prediction of "-0.058" indicates that the stock reduces by 0.058 points and the accuracy of the prediction in this case is 98.7%.

E. Calculating the performance of the model

In this stage, the machine learning model is then tested with the testing data which is a very essential step as it determines whether the training samples in the previous stage was processed properly and whether the results of the test data from the machine learning model can be used to check for real-time news articles and get the desired results.

F. Accuracy of the ML model

After testing the machine learning[6] on the test data, it is now required to check the accuracy of the model. How do we determine the accuracy of the machine learning model?
The model checks whether the predicted results of the dataset matches the actual results and then it takes an overall average percentage and outputs a percentage.
It is found that using the DenseLayers network consisting of 3 layers,
i) input layer: consisting of 128 nodes
ii) hidden layer: consisting of 128 nodes
iii) output layer: consisting of 1 node,
an accuracy of **55.45%** for the dataset used.

III. RESULTS

In general, an accuracy of 55% was achieved with a high of 61% and a low of 45%.

Some snapshots are as follows:

```
In [11]: df['vader'].plot(kind='hist')
         plt.xlabel("Polarity")
         plt.ylabel("Frequency")
         plt.show()
```

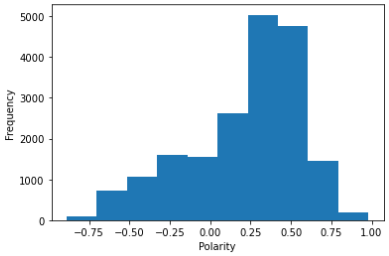

Fig. 1. Bar graph showing the relation between polarity and frequency

```
In [10]: #Evaluate
         results = (model.evaluate(X_test,y_test, batch_size=10,verbose=False))
         print(results)
         predictions = model.predict(X_test)
         plt.scatter(y_test, predictions)
         plt.xlabel("True Values")
         plt.ylabel("Predictions")

         [0.9848762486512954, 1.5326259, 1.0, 0.0]

Out[10]: Text(0, 0.5, 'Predictions')
```

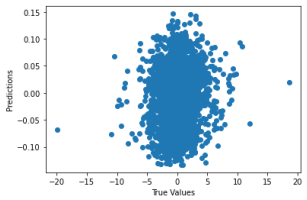

Fig 2: Scatter plot showing the prediction with relation to real values

```
date,stock,vader,secscore,assoc,perc,percword,sector,index,news
2012-09-09,Larsen & Toubro,0,-0.49,0.0,0.2,'positive','Construction','LT','news'
2012-09-09,BPCL,0,0.36,0.0,-0.7,'negative','Energy','BPCL','news'
2012-09-09,GAIL,0,0.36,0.0,1.0,'positive','Energy','GAIL','news'
2012-09-09,IOC,0,0.36,0.0,0.88,'positive','Energy','IOC','news'
2012-09-09,ONGC,0,0.36,0.0,0.49,'positive','Energy','ONGC','news'
2012-09-09,Reliance Industries,0,0.36,0.0,-0.99,'negative','Energy','RELIANCE','news'
2012-09-09,NTPC Limited,0,-0.12,0.0,0.5,'positive','Power','NTPC','news'
2012-09-09,PowerGrid Corporation of India,0,-0.12,0.0,0.49,'positive','Power','POWERGRID','news'
2013-05-19,Larsen & Toubro,0,0.3,0.0,-0.06,'negative','Construction','LT','news'
2013-05-19,BPCL,0,0.27,0.13,-2.27,'negative','Energy','BPCL','news'
2013-05-19,GAIL,0,0.27,0.13,-1.38,'negative','Energy','GAIL','news'
2013-05-19,IOC,0.67,0.27,0.13,-1.57,'negative','Energy','IOC','news'
```

Fig 3: CSV file of the final readings

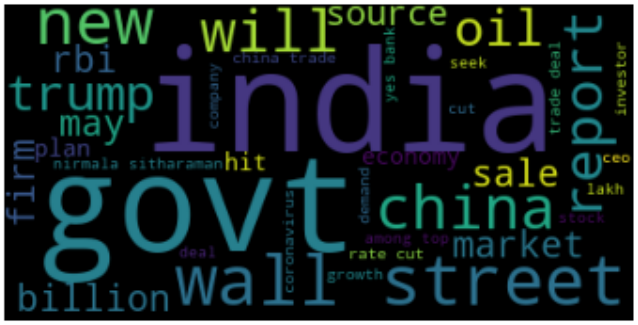

Fig 4: Word cloud of the most common words in the dataset

CONCLUSION

In this project, we have significant proof that there is a correlation between the price of shares of stock and the daily news associated with it. Through the machine learning model, we were able to observe that there is in fact an impact through these news articles. Though not much, it is helpful for an enthusiast who is deeply passionate about investing in the stock market

This project was carried out on NIFTY50 company dataset and the corresponding news dataset for each of those 50 companies. The prediction of the model that we were able to achieve was roughly around 55% with the highest being 61% and the lowest being 45%.

ACKNOWLEDGEMENT

We would like to thank our college for giving us this opportunity. We would also like to thank our guide, the HOD for her guidance.

We would like to thank our friends and family without whom we would not have been able to complete this project.

REFERENCES

[1] Dev Shah, Haruna Shah, Farhana Zulkernine, “Predicting the effects of news sentiments on the stock market,” 2018 IEEE conference on Big Data(Big Data), ISBN:978-1-5386-5035-6/18

[2] Yasef Kaya, M. Elif Karsligil, “Stock price prediction using financial news articles”, 2010 IEEE , ISBN: 978-1-4244-6928-4/10.

[3] HD Huynh, LM Dang, D Duong, “A New model for stock price movements prediction using Deep Neural Network”, SoICT, 2017, pp.57-62: ACM

[4] Stock market prediction using daily news articles: Yashwanth Singh Patel, Supriyo Mandal, IIT Patna, 2017

[5] Cicil Fonseka, Liwan Liyanage, “A data mining algorithm to analyse stock market data using lagged correlation”, 2008 IEEE, ISBN: 978-1-4244-4/08.

[6] Bhargav Hegde, Dayananda P, Mahesh Hegde, Chetan C, “ Deep Learning Technique for Detecting NSCLC”, International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-3, September 2019, pp. 7841-7843. DOI: 10.35940/ijrte.C6540.098319

[7] Kalyani Joshi, Prof.Bharati, Prof. Jyothi Rao, “Stock trend prediction using news sentiment analysis”, IJCSIT VOL.8 No.3 June 2016