Analyzing stock trend using daily news articles

Using Natural Language processing and Machine Learning

Shravan Bhat
Computer Science, K.S Institute of Technology, Bangalore, Karnataka, India, shravanbhat98@gmail.com

Siddhanth M
Computer Science, K.S Institute of Technology, Bangalore, Karnataka, India, siddhanth8@gmail.com

Sampath Kumar
Computer Science, K.S Institute of Technology, Bangalore, Karnataka, India,
sampath.kulkarni7@gmail.com

Dr. Rekha B Venkatapur
Head of Department, Computer Science, K.S Institute of Technology, Bangalore, Karnataka, India,
rekbhabvenkatapur@ksit

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 VG Siddhartha's demise, founder of Cafe Coffee Day (CCD), affected the share prices of CCD as it **dropped by 60%**. 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.

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
This will help investors and stock market traders to have an informed decision on what to invest in. The main problems that we are trying to solve are to improve the accuracy of the prediction model, and to make the model adaptive to more than one dataset.

KEYWORDS

Machine learning, Preprocessing, Natural Language processing, Stock market prediction, Neural Networks, Correlation

1. 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.

2. METHODOLOGY

The project is broken into 6 parts:

PART I: Data collection and sentiment analysis on the collected data.

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.

2.1 Data Collection

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.

Determining the polarity of the news article

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.

2.2 Developing 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 model

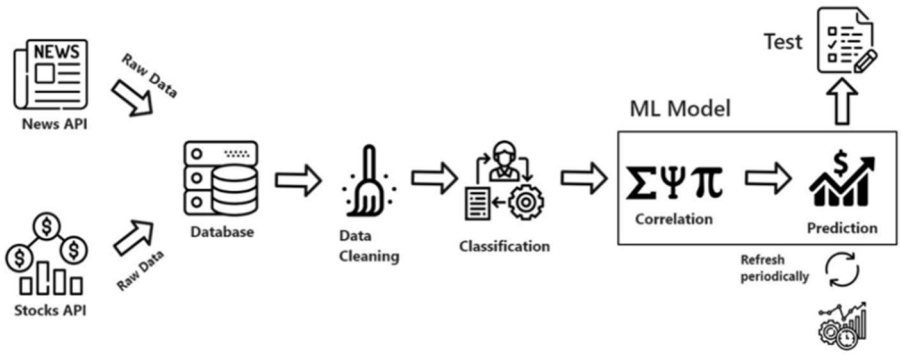

Figure 1: Dataflow of the system

2.3 Training the ML model with data

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, since we have used 79,000*500 = 39500000 samples

fit() is the method used in Tensorflow to invoke the training process

2.4 Calculating performance of the model

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

53/53 [==============================] - 0s 2ms/step - loss: 0.9880 - accuracy
[0.9879634976387024, 0.0]
```

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%

.

2.5 Testing the ML 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.

The code for testing the machine learning model is:

```
input_val = ([[-0.232,-0.2,-0.23]]) input_val = np.asarray(input_val)
prediction = model.predict(input_val) print("Input: ",input_val)
y_0 = prediction[0][0] print("Prediction:", y_0)
predictions = model.predict(X_test)
print(model.evaluate(X_test,y_test)) #percentage drop in stock price
print(X_test[1:2],predictions)
```

2.6 Check accuracy of the model

After testing the machine learning 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 (as we require only the percentage change of the stock),

we are able to achieve 55.45% accuracy for the dataset used.

```
15291/15291 [==============================] - 1s 35us/sample - loss: 0.9844 - mean_absolute_error: 1.5470 - categorical
_accuracy: 1.0000 - accuracy: 0.0000e+00
Epoch 497/500
15291/15291 [==============================] - 0s 30us/sample - loss: 0.9849 - mean_absolute_error: 1.5471 - categorical
_accuracy: 1.0000 - accuracy: 0.0000e+00
Epoch 498/500
15291/15291 [==============================] - 1s 34us/sample - loss: 0.9851 - mean_absolute_error: 1.5471 - categorical
_accuracy: 1.0000 - accuracy: 0.0000e+00
Epoch 499/500
15291/15291 [==============================] - 0s 31us/sample - loss: 0.9850 - mean_absolute_error: 1.5475 - categorical
_accuracy: 1.0000 - accuracy: 0.0000e+00
Epoch 500/500
15291/15291 [==============================] - 1s 33us/sample - loss: 0.9843 - mean_absolute_error: 1.5471 - categorical
_accuracy: 1.0000 - accuracy: 0.0000e+00
Accuracy  = 55.45
Confusion =
[[1212  751]
 [ 952  908]]
Took  245.52210211753845
```

Figure 2: Accuracy output of the model

On a per week basis by direct positive score to positive stock market score, we found that it ranges from 47% to 61%. Without the mean sector scores, the average accuracy was reduced by 2%. If we could include more parameters and data from the stock market and other global indices, we could enhance this accuracy to a higher value.

3. RESULTS

```
headline,url,date1,date2
Dance reviews,https://www.deccanherald.com/content/216090/dance-reviews.html,2012-01-01,2012-01-01
Govt to charge for spectrum beyond 4.4 MHz from new players,https://www.deccanherald.com/content/215938/govt-
Stock market hopes for better days in 2012,https://www.deccanherald.com/content/215939/stock-market-hopes-bet
It's happy new year for hiring; over 5 lakh new jobs in 2012,https://www.deccanherald.com/content/215948/its-
SEBI committee examining e-IPO proposal,https://www.deccanherald.com/content/215981/sebi-committee-examining-
Toyota Kirloskar Motor sales jump 82 pc in 2011,https://www.deccanherald.com/content/215982/toyota-kirloskar-
```

Figure 3: This is the input which we downloaded from news sources

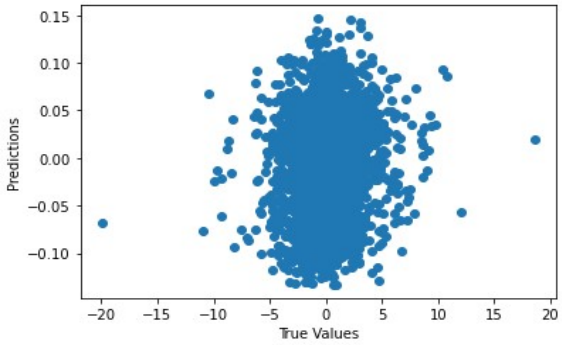

Figure 4: This represents all the prediction values in binary range

```
1   date,stock,vader,secscore,assoc,perc,percword,sector,index,news
2   2012-09-09,Larsen & Toubro,0,-0.49,0.0,0.2,'positive','Construction','LT','news'
3   2012-09-09,BPCL,0,0.36,0.0,-0.7,'negative','Energy','BPCL','news'
4   2012-09-09,GAIL,0,0.36,0.0,1.0,'positive','Energy','GAIL','news'
5   2012-09-09,IOC,0,0.36,0.0,0.88,'positive','Energy','IOC','news'
6   2012-09-09,ONGC,0,0.36,0.0,0.49,'positive','Energy','ONGC','news'
7   2012-09-09,Reliance Industries,0,0.36,0.0,-0.99,'negative','Energy','RELIANCE','news'
8   2012-09-09,NTPC Limited,0,-0.12,0.0,0.5,'positive','Power','NTPC','news'
9   2012-09-09,PowerGrid Corporation of India,0,-0.12,0.0,0.49,'positive','Power','POWERGRID','news'
10  2013-05-19,Larsen & Toubro,0,0.3,0.0,-0.06,'negative','Construction','LT','news'
11  2013-05-19,BPCL,0,0.27,0.13,-2.27,'negative','Energy','BPCL','news'
12  2013-05-19,GAIL,0,0.27,0.13,-1.38,'negative','Energy','GAIL','news'
13  2013-05-19,IOC,0.67,0.27,0.13,-1.57,'negative','Energy','IOC','news'
```

Figure 5: This is a snapshot of the dataset that is fed into the machine learning model from which only certain columns are selected for prediction. Those columns are stock_name, sector_score, perc_change

ILLUSTRATION : Developing the Neural network

```
model = tf.keras.Sequential()

model.add(tf.keras.layers.Dense(64, activation=activation,input_shape=(X_train.shape[1],)))

model.add(tf.keras.layers.Dense(132, activation=selu))

model.add(tf.keras.layers.Dense(64, activation=selu))

model.add(tf.keras.layers.Dense(1))
```

model.compile(loss=squared_hinge, optimizer=adamax,metrics=[metrics.mae, metrics.categorical_accuracy,metrics.accuracy])

4. CONCLUSION AND FUTURE WORK

In this project, we have significant proof that there is a strange 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%.

This model can be used in the future by including more parameters from the global indices, such as the forex markets, other stock market trading platforms etc.

The dataset here has been exclusively financial news articles and hence. a future upgrade would be to include other forms of news as well.

NOTE:

The dataset used in the month of February and March 2020 is highly varied because of the COVID-19 pandemic. Hence the results for these months are very low and hence have had an impact on the model's accuracy as well.

ACKNOWLEDGMENTS

We would like to thank our college management,for their enormous support.
We would like to thank our project guide, Dr. Rekha B Venkatapur, for mentoring and the constant input that we got while developing the project.
We would also like to sincerely thank our parents and friends for the constant support and motivation.

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