

# Analysis of the Effect of News Sentiment on Stock Market Prices through Event Embedding

Sashank Sridhar  
College of Engineering Guindy,  
Anna University, Chennai, India  
Email:  
sashank.ssridhar@gmail.com

Sowmya Sanagavarapu  
College of Engineering Guindy,  
Anna University, Chennai, India  
Email:  
sowmya.ssanagavarapu@gmail.com

**Abstract**—Stock market price prediction models have remained a prominent challenge for the investors owing to their volatile nature. The impact of macroeconomic events such as news headlines is studied here using a standard dataset with closing stock price rates for a chosen period by performing sentiment analysis using a Random Forest classifier. A Bi-LSTM time-series forecasting model is constructed to predict the stock prices by using the polarity of the news headlines. It is observed that Random Forest Classifiers predict the polarity of news articles with an accuracy of 84.92%.

## I. INTRODUCTION

NEWS plays a significant role in the investment world as it provides information to the investors to make decisions in the stock markets. It is capable of shaping and influencing the emotions and opinions of people, driving the decision to buy or sell in markets. Recently, an example of this was the world markets crashing in March 2020 [1] during the COVID-19 global pandemic. Due to the imposition of nationwide lockdowns forcing businesses to close down and stop their ongoing activities, investors were faced with uncertainty, leading to the markets across the world crashing in March 2020.

Analysis of media sentiments enables performing text analysis to determine its opinion and the subjectivity. In his Economic Research paper, Samuel P. Fraiberger of the World Bank [2] has shown that news sentiment acts as an important predictor of daily stock returns in stock markets. Sentiment analysis [3], or opinion mining, is a Natural Language Processing (NLP) technique to determine the polarity of the text sentiment (positive, negative, or even neutral). Machine learning architectures such as Support Vector machines, Boosting and Bagging algorithms and Random Forests perform this analysis by assigning sentiment scores to the categories within a phrase in a sentence to determine polarity.

Random Forests is a Machine Learning Algorithm that is built using multiple decision trees merged together for an accurate and stable prediction. This algorithm also adds randomness to the data for enhancing its performance, while training using the data bagging algorithm.

Stock market price prediction models have helped in the determination of asset investments for maximizing individual profits. The complexity of the time-series

forecasting tasks are handled by Bidirectional Long Short Term Memory (Bi-LSTM) [4].

In this paper, sentiment analysis is performed on the varying sets of news headlines dataset collected for each day to analyze the polarity of the data as positive or negative. The positive class refers to the headlines which led to the increase in the stock price the next day and the negative class indicated the drop in the stock price. Using a random forest classifier, the data was studied and the results were analyzed for the impact of polarity prediction on stock price forecasting using Bi-LSTMs.

The rest of the paper is organized as follows. Section II gives the summary of some of the best works in stock market price prediction modelling. The design of the system for market price prediction using news and the sentiment analysis model for prediction of rate change is given in Section III. The implementation details of the models are given in Section IV. The results obtained from the implemented price prediction system is presented in Section V. The summarization of the project and the proposed future work is given in Section VI.

## II. RELATED WORKS

In this section, some of the recent works in stock market price prediction using sentiment analysis have been summarized.

Stock movement prediction model using dilated causal convolutions and transformer modelling was discussed by Daiya and Lin [5]. They extracted features from the data to feed into a multi-head self-attention model by considering financial indicators and news data. A basic reinforcement learning policy and reward function to match with their performance was used in their model. By implementing multimodal learning, the model aimed to maximize forecasted profits through asset allocation based on the prediction modelling.

Case studies dealing with the effect of public sentiment in stock market price predictions using big data analysis have also been published. Bourezk et al [6] used machine learning algorithms to analyze the relationship between the general public view regarding a stock and its evolution within the Moroccan Stock Exchange. Malawana and Rathnayaka [7] performed sentiment analysis on market related announcements in the news to extract positive, negative and

neutral opinions. Using Naive Bayes and Linear Regression models for this, they performed detections of sentiment class within a Big Data distributed Environment.

### III. SYSTEM DESIGN

The overall architecture for the prediction of stock prices and the rate of movement is given in Figure 1. Multitask architecture comprises a sentiment analysis module which determines the direction of movement of stocks from news headlines.

#### I.A. Data Set Description

The dataset used is collected from [8] and it consists of stock price data for Dow Jones Industrial Average (DJIA) and corresponding news article headlines regarding the stock index from the period between 2008-08-08 to 2016-07-01. The headlines for each day are annotated as either 0 when the close price decreased compared to the previous day and 1 when the close price stayed the same or rose compared to the previous day.

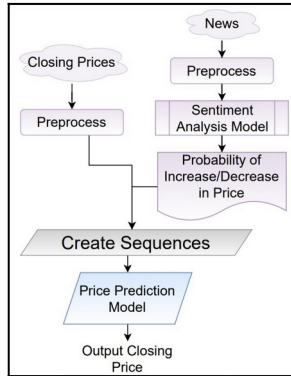


Fig. 1. Overall Multitask Architecture for Prediction

#### I.B. Preprocessing the Price Dataset for Price Prediction and Rate of Change Analysis

The price dataset is first normalized to ensure all the values are in the range (0,1) that helps the model to converge at the local minima at a faster rate. The dataset is converted into input sequences of length  $n$  and the corresponding output price is the  $(n+1)$ th price in the sequence. This creates a sliding window of fixed length whose output corresponds to the forecasted prices.

#### I.C. Preprocessing News Headlines for Sentiment Analysis

Data cleaning is performed and each word in a sentence is converted into a vector form to be modelled by assigning a Term frequency - inverse document frequency (Tf-IDF) score [9].

#### I.D. Multitask Learning

Multitask learning [10] aims to learn individual sub-tasks separately and use those learnings inductively to solve a main task by identifying the dependence between the tasks. Separate multitask models are built to predict the rate of change of stock prices and to predict the actual stock price itself. The subtasks involve modelling the prices and

identifying the sentiment of the news headlines and these subtasks act as Level-0 models.

#### I.E. Event Embedding with Sentiment Analysis

The vectorized headlines are fed to different machine learning models in order to predict their positive sentiment corresponding to the close prices being steady or increasing and negative sentiment corresponding to the close prices decreasing. The obtained probabilities of sentiment from the machine learning models act as events [11] that contribute to the price prediction model. The events are converted into sliding windows to correspond to the sliding window of prices and are given as input to the overall price prediction multitask model. Figure 2 shows how event sequences are generated using the polarity derived from the sentiment analysis models.

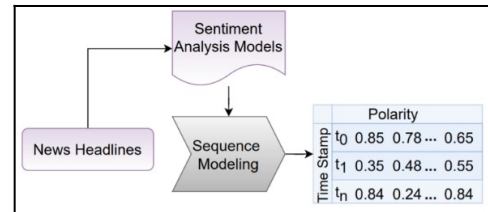


Fig. 2. Event Embedding of Sentiment Polarity

#### I.F. Price Modelling

Sequences of stock prices are proposed to be modelled using Long Short-Term Memory (LSTM) models and the modified Bi-Directional LSTM (Bi-LSTM) models. LSTM models make use of gates to determine if the data at a particular node should be retained or not from the cell state to map future and past interrelations between the data. Bi-LSTMs use the LSTM cells to map the dependencies between sequences in both the forward and reverse directions.

#### I.G. Layer-1 Meta Classifier

Once the subtask models are built the output of the models are fed to a Layer-1 Meta-Classifer that relates both the tasks. The meta-classifier used is an Artificial Neural Network (ANN). The outputs of the price prediction sub-model and the sentiment analysis model are given as inputs to the ANN which then predicts the output closing price of the given stock index.

## IV. SYSTEM IMPLEMENTATION

The implementation details of the system are given in this section.

### A. Dataset Split

#### 1) Sentiment Analysis Model

The sentiment analysis model has top 25 news headlines for each day from 2008-08-08 to 2016-07-01. The dataset was divided in a train-test ratio of 80: 20 as seen in Table I.

#### 2) Price Prediction Model

The dataset was divided into a train-test ratio of 80:20 with the training set consisting of prices for 1863 days and the testing set consisting of prices for 378 days. Prices are

transformed to form a sliding window with an input sequence length of 50 and an output sequence length of 1.

### B. Sentiment Analysis

The Tf-IDF score is calculated for news headlines using sklearn's Tf-IDF Vectorizer. Each day's headlines are annotated with labels of 0 or 1 corresponding to a decrease or increase in prices on that day. Different machine learning algorithms are implemented with their corresponding parameters as seen in Table II.

### C. Price Prediction

The input closing prices is first normalized using sklearn's MinMaxScaler which ensures that all prices remain between 0 and 1. The model comprises 3 Level-0 models corresponding to modelling of prices, negative sentiment and positive sentiment probabilities of the news headlines. The sub-models that learn the positive and negative sentiment of the headlines. The outputs of all the three sub-models are concatenated and fed to a Level-1 meta-classifier with ReLu Activation. The prediction model is trained for 1000 epochs in batches of size 8 with RMSProp as the optimizer and MSE as the loss function

## V. RESULTS AND ANALYSIS

The results obtained from the stock price prediction model and sentiment analysis models are presented in this section along with their analysis.

### A. Evaluation metrics for the sentiment analysis model using Machine Learning algorithms

The standard dataset contains day-by-day news headlines along with whether the stock market price increased (positive class) or decreased/remained the same (negative class) the next day for sentiment analysis classification of

the text as positive and negative class. It is observed from Table III, that the Random Forest (RF) algorithm has outperformed the other algorithms with the no. of headlines at 25. This RF model [12] is composed of a number of decision tree classifiers that help to identify the important gestures from the dataset for its high performance.

### B. Evaluation metrics for the Random Forest sentiment analysis model

Sentiment analysis was carried out by using RF classifier on a day's news headlines to predict the increase or decrease in the stock price for the next day. The random forest-based machine learning algorithm was tested with multiple numbers of headlines chosen on each run to identify and record its optimal performance. The positive class or class-1 refers to the news headlines that predicted the increase in stock price the next day and negative class or class-0 when the closing price of the stock remained the same or decreased. From Table IV, it's observed that it performed best with 25 headlines, reaching a performance accuracy of 84.92. The confusion matrix for the RF based sentiment analysis model is given in Figure 3.

### C. Calculation of sentiment score for positive and negative sentiment

The sentiment score is calculated for every news headline from the dataset. In a headline with a positive sentiment, count for each word in both the positive counter and the total words in the dataset counter; likewise, for each word in a negative sentiment headline, count for that word in both the negative counter and the total words counter is increased.

The most commonly occurring words belonging to the news headlines from the positive and negative sentiment are extracted and visualized in Figure 4. The headlines that referred to new releases seemed to have one of the highest impacts on the stock market prices. News related to hacking, sanctions and scandals have had the highest negative impact resulting in the fall of the stock prices.

### D. Prediction of Stock Prices with News Sentiment Analysis

The daily stock prices are plotted in the graph in Figure 5 and compared with the predicted values from the Bi-LSTM model trained with news headlines for performance analysis. It is observed that the prediction results with top 25 news headlines shows that the model is able to predict the positive

TABLE II.  
DATASET SPLIT FOR SENTIMENT ANALYSIS MODEL

Data Type	Train Set	Test Set
Number of Days	1863	378
Number of News Headlines	50301	10206
Headlines with Positive Sentiment	26865	5184
Headlines with Negative Sentiment	23436	5022

TABLE II.  
PARAMETERS OF SENTIMENT ANALYSIS CLASSIFIERS

Parameter	Value
Random Forest	n_estimators=200, criterion='entropy'
Support Vector Machine	kernel='linear'
Adaboost Classifier	n_estimators=200
Bagging Classifier	base_estimator=Support Vector Classifier, n_estimators=10, random_state=0
Decision Tree Classifier	random_state=0

TABLE III.  
PARAMETERS OF SENTIMENT ANALYSIS CLASSIFIERS

Model	No. of headlines					
	1	5	10	15	20	25
RF	84.1	83.6	84.7	83.6	81.7	84.9
SVM	82.8	83.9	83.3	83.6	84.6	84.6
Adaboost	63.7	72.2	71.9	74.6	73.5	74.3
Bagging	80.6	81.4	81.4	81.4	81.4	81.9
Decision Tree	82.0	81.7	77.5	81.5	77.5	80.68

or negative trend change on the next day with a high value of accuracy through NLP techniques.

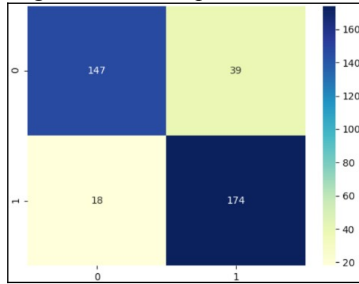


Fig. 3. Confusion Matrix obtained from the Random Forest Model

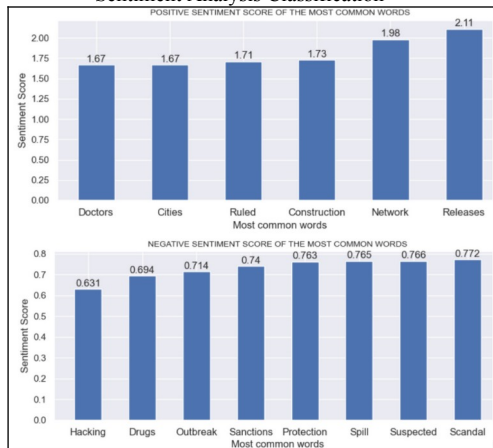


Fig. 4. Most Common Words from the Dataset with a) High Positive Score and b) High Negative Score

IV. CONCLUSION

The work explores the effect of public sentiment through news headlines on stock market prices. To accomplish that, a Random Forest classifier-based sentiment analysis model is constructed using day-by-day news headlines dataset along with the variation of Close Stock price. The sentiment analysis model identified the headlines associated with positive and negative sentiment for further analysis. The constructed Multitask Bi-LSTM based stock price prediction model was used for predicting the close price rates with news headlines dataset. A deep neural network architecture-based stock price prediction model is to be constructed to experiment with using trained weights from the sentiment analysis performed for optimizing the learning weights of the model further with attention-based deep neural

TABLE IV.  
PERFORMANCE OF THE RANDOM FOREST SENTIMENT ANALYSIS MODEL

Evaluation Metric	Performance of the model in %
Accuracy	84.92
Precision	85.0
Recall	85.0
F1-Score	85.0

architectures. This model would be analyzed to calculate the rate of change of price with chosen top-n headlines for positive and negative sentiment in the news articles.

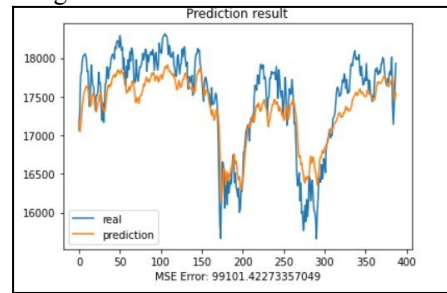


Fig. 5. Real and Prediction Value comparison for the trained Bi-LSTM Multitask Model with News Headlines Data

REFERENCES

- J.-J. Ohana, S. Ohana, E. Benhamou, D. Saltiel, and B. Guez, "Explainable AI Models of Stock Crashes: A Machine-Learning Explanation of the Covid March 2020 Equity Meltdown," SSRN Electronic Journal, 2021, doi: <http://dx.doi.org/10.2139/ssrn.3809308>.
- S. P. Fraiberger, D. Lee, D. Puy, and R. Ranci re, "Media Sentiment and International Asset Prices," NBER Working Papers 25353, National Bureau of Economic Research, Inc., 2018.
- M. Skuza and A. Romanowski, "Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction," in 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1349–1354, 2015, doi: 10.15439/2015F230.
- D. Ruta, L. Cen and Q. H. Vu, "Deep Bi-Directional LSTM Networks for Device Workload Forecasting," in 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 115-118, 2020, doi: 10.15439/2020F213.
- D. Daiya and C. Lin, "Stock Movement Prediction and Portfolio Management via Multimodal Learning with Transformer," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3305–3309, 2021, doi: 10.1109/ICASSP39728.2021.9414893.
- H. Bourezk, A. Raji, N. Acha, and H. Barka, "Analyzing Moroccan Stock Market using Machine Learning and Sentiment Analysis," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5, 2020, doi: 10.1109/IRASET48871.2020.9092304.
- M. V. D. H. P. Malawana and R. M. K. T. Rathnayaka, "The Public Sentiment analysis within Big data Distributed system for Stock market prediction– A case study on Colombo Stock Exchange," in 2020 5th International Conference on Information Technology Research (ICITR), pp. 1–6, 2020, doi: 10.1109/ICITR51448.2020.9310871.
- J. Sun, "Daily News for Stock Market Prediction, Version 1," kaggle.com, 2016. <https://www.kaggle.com/aaron7sun/stocknews> (accessed May 23, 2021).
- J. A. Reyes-Ortiz, M. Bravo, and H. Pablo, "Web Services Ontology Population through Text Classification," in 2016 Federated Conference on Computer Science and Information Systems, pp. 491–495, 2016, doi: 10.15439/2016F332.
- A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," in 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 179–183, 2020, doi: 10.15439/2020F20.
- P. Maciag, "Efficient Discovery of Top-K Sequential Patterns in Event-Based Spatio-Temporal Data," in 2018 Federated Conference on Computer Science and Information Systems, pp. 47–56, 2018, doi: 10.15439/2018F19.
- J. Lind n, S. Forsstr m, and T. Zhang, "Evaluating Combinations of Classification Algorithms and Paragraph Vectors for News Article Classification," in 2018 Federated Conference on Computer Science and Information Systems, pp. 489–495, 2018, doi: 10.15439/2018F110.