Stock Price Prediction using news article and numeric data

KEVIN JADIYA | MTECH-ML | 202111010, DAIICT, India
MAULIK SARVAIYA | MTECH-ML | 202111025, DAIICT, India
RAHUL VANSH | MTECH-ML | 202111035, DAIICT, India

Abstract

When the stock market is open, financial news stories cover economic developments that have an impact on the stock market. Here, the direction of close price movement is predicted using information from text features and price features.

Additional Key Words and Phrases: Word Cloud, Entity Tagger, Sentiment Analysis, Issue Detection

ACM Reference Format:
Kevin Jadiya | Mtech-ML | 202111010, Maulik Sarvaiya | Mtech-ML | 202111025, and Rahul Vansh | Mtech-ML | 202111035. 2023. Stock Price Prediction using news article and numeric data. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Financial news stories cover economic developments that have an impact on the stock market. When the stock market is open, news pieces are released that have an impact on how the stock market indices move. There is an implied connection between various economic developments that ultimately affect stock market movement. This connection between news reports and the direction of the stock market close price movement can be helpful in predicting the direction of the stock market. Deep learning-based models employ price data or text features as input to forecast the close price movement in a specific direction.

This study also suggests a parallel CNN model that predicts the direction of close price change by fusing text representation with price representation.[2].

You can find our code[4] at github https://github.com/rahulvansh66/share_market_direction_prediction.

2 IMPLEMENTATION

For implementation, we took reference from [3] code. We mainly focused on deep learning models since it's giving better results than classical machine learning models. This problem can be seen as a classification problem. The label is set to 1 if the closing price of the current trading day is higher than the closing price of the previous trading day, and 0 otherwise.

From experiments, it's been observed that CNN with FinBERT[1] for text representation gave the best results. Upcoming sections will contain more detailed information about these.

Manuscript submitted to ACM

2.1 Price only

For all of our experiments, we use historical NIFTY 50, NIFTY Next 50, and NIFTY Bank prices, as well as financial news articles from 2010 to 2021. There are 2978 trading days in each index. The following technical indicators are also calculated for all indices: Average Directional Index (ADX), Moving Average Convergence Divergence (MACD), Momentum (MOM), Average True Range (ATR), Relative Strength Index (RSI), Stochastic Oscillator (STOCH), William's %R (WILLR), Bollinger Bands (BBANDS), Exponential Moving Average (EMA), and Simple Moving Average (SMA).We also use archive news articles from the Economic Times from 2010 to 2021 for the news article dataset. The headline, publication date, and text of the news stories are all included. For all of our experiments, we choose news article headlines between 9 AM and 4 PM IST because this is when the market is open every day. To eliminate the news that isn't about finance, we use a rule-based classifier.

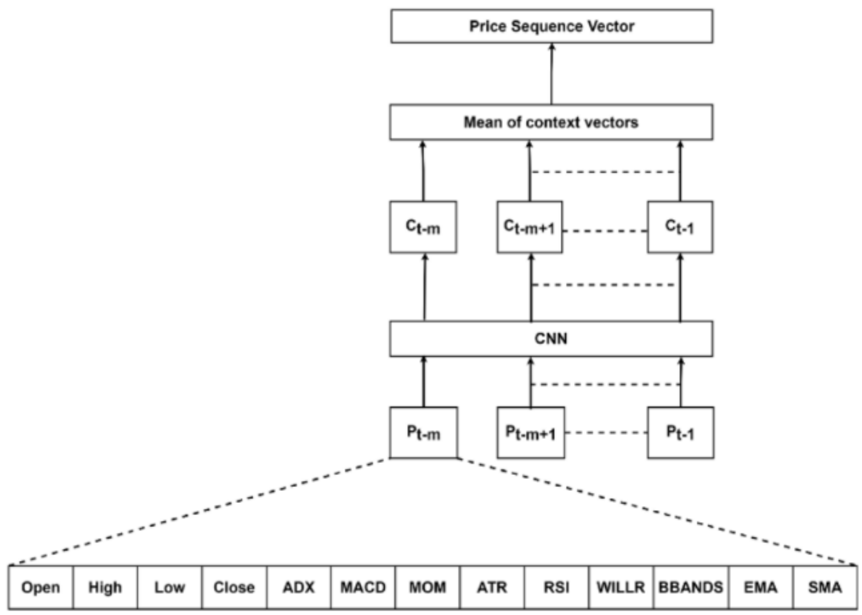

Fig. 1. Price only given as input[2]

2.2 Text only

Here, we take the news article's title as input and convert it into the vector using FinBERT. Then we use CLS token v_{CLS} of each news article which represents the entire sentence. The news vectors are then pooled together to form the news sequence vector. CNN is applied to the sequence of news title vectors by the news sequence encoder. The context vectors that capture the local relationships between the news articles will be the CNN layer's output. These context vectors are identified by the notation [c1, c2,...., cN]. The context vectors are averaged, leading to the exchange of information that each context vector has encoded. The averaged vector is the news sequence vector representing the sequence of news article titles[2].

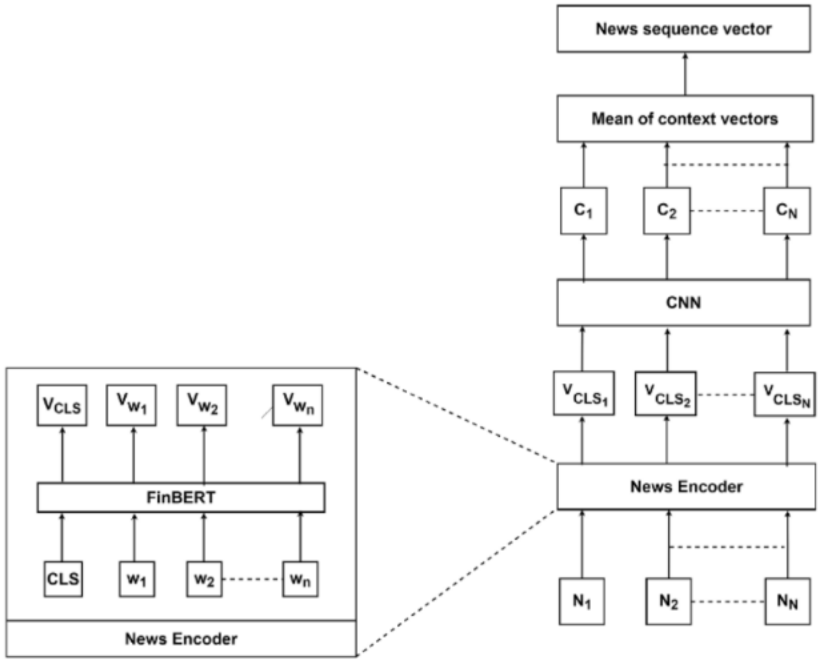

Fig. 2. News articles will be given as input[2]

2.3 Combining Text and Price

Now let's look at how we can combine news articles and prices and use the internal relationship to predict the direction of stock price. As shown in the below figure, the architecture consists of the following components; News Encoder, News Sequence Encoder, Prices Sequence Encoder, and Prediction Layer. The latter figure shows the algorithm used for training and testing the model.

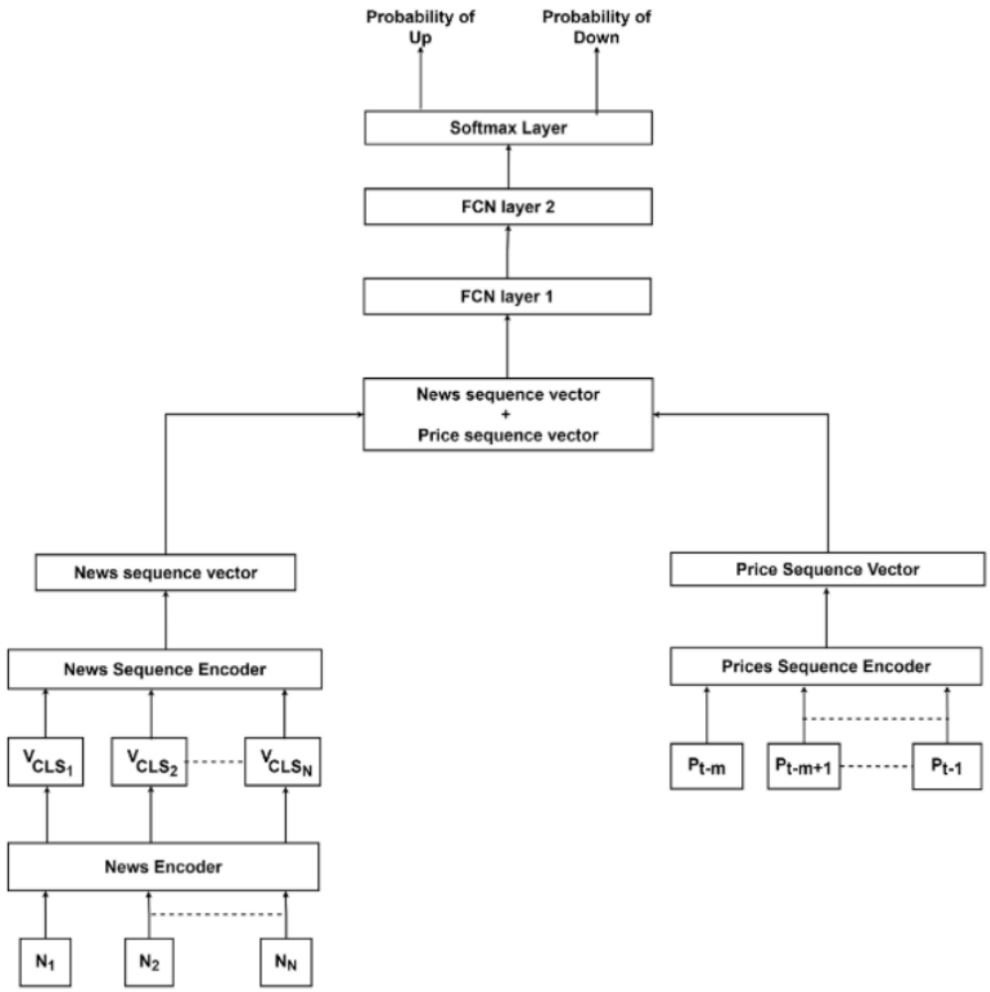

Fig. 3. Price and news articles will be given as input[2]

```
Input: News Articles [N_1, N_2, ....., N_N] and Historical Prices
        [p_1, p_2, ...., p_m]
Output: Direction of close price movement
train_years = [2010, 2011, 2012, 2013, 2014, 2015, 2016]
test_years = [2017, 2018, 2019, 2020, 2021]
parallel_cnn = Initialize Classifier
epochs = Number of epochs
scaler = Initialize Z normalization scaler
for each test_year in test_years do
    for each epoch in epochs do
        Calculate technical indicators from prices
        Fit scaler to normalize the prices and technical indicators
        for each each day in train_years do
            Select news articles between 9 AM to 4 PM
            Select label assigned to the day
            Represent each news articles using text representation
              scheme
            Use prices and news representation to train parallel_cnn
        end
    end
    for each each day in test_year do
        Select news articles between 9 AM to 4 PM
        Represent each news articles using text representation
          scheme
        Apply scaler to normalize the prices and technical
          indicators
        Use the news representation, prices and technical
          indicators to predict the label using parallel_cnn
    end
    Evaluate parallel_cnn using predicted labels and test labels
    Add test_year to train_year
end
```

Fig. 4. Pseudo-code of Combining Text and Price only approach[2]

3 RESULTS

The results are shown in the below table. We can clearly see that we are getting better results when we train the model for more years, also if we consider the past Fourteen days rather than just seven days, we are getting remarkable improvements.

Table 1. Results using text and price only approach

Approach	Training years	Testing year	past days considered	Accuracy	Lag	Cross-correlation	ROC-AUC-score
Price only	2010 to 2016	2017	7	0.47899	32	0.16155	0.48780
Text only	2010 to 2016	2017	7	0.50403	1	0.15232	0.51658
Price and Text only	2010 to 2016	2017	7	0.6099	0	0.18291	0.56649
Price and Text only	2010 to 2017	2018	7	0.67782	0	0.35551	0.67786
Price and Text only	2010 to 2017	2018	14	0.70258	0	0.43239	0.70961

4 INDIVIDUAL CONTRIBUTION

Kevin and Rahul read the paper suggested by you while Maulik was learning stock market terms. Later Rahul started learning transformers while Maulik was learning parallel CNN, which helped us to understand the state of the art of model proposed in the paper. While Kevin and Rahul continued reading the paper, Maulik started learning and implementing code. Then, we shared our learning with each other so that each one of us had a good understanding of paper and code at the same time. We also discussed which observations to take and what kinds of chances we should try. Based on that, we divide the task of running code with different modifications. Among many experiments, we mentioned only good results that worked well. Results of 'Price only' and 'Text only' were taken by Kevin. 'Price and text only' for the past seven days and 2017 as a test taken by Rahul. 'Price and text only' for the past 7 and further for the past 14 days too, as you mentioned that you want us to consider more past days' news to predict the direction of the stock market, we also consider 2018 as a test while used 2010 to 2017 as training, this observation has been taken by Maulik.

ACKNOWLEDGMENTS

We would like to thank Prof. Prasenjit Majumder and our TA, Surupendu Gangopadhyay sir, for guiding us throughout developing this project.

REFERENCES

[1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *CoRR* abs/1908.10063 (2019). arXiv:1908.10063 http://arxiv.org/abs/1908.10063

[2] Surupendu Gangopadhyay and Prasenjit Majumder. 2023. Text representation for direction prediction of share market. *Expert Systems with Applications* 211 (2023), 118472. https://doi.org/10.1016/j.eswa.2022.118472

[3] surupendu. 2022. *share market direction prediction.* Retrieved December 2, 2022 from https://github.com/surupendu/share_market_direction_prediction

[4] Rahul Vansh. 2022. *share market direction prediction.* https://github.com/rahulvansh66/share_market_direction_prediction