

Effect of Daily News on Stock Fluctuation An NLP-Based Approach

Shouvik Dasgupta

School of Engineering Sciences & Technology
Jamia Hamdard University
India
shouvikdasgupta3125@gmail.com

Abstract – Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. Stock fluctuation prediction is done to determine whether the current stock value will rise, fall or remain the same. One of such ways of predicting the fluctuation can be the news headlines of that particular day. Using an NLP based approach and a baseline classification model, we have tried to determine whether daily news can be used to predict stock fluctuation.

Keywords – Artificial neural network, Bag of words, Bert transformers, TF-IDF, Machine Learning, Natural language Processing, Stock Prediction, Word2Vec.

I. INTRODUCTION

Research in stock market prediction has come a long way. Numerous approaches have been tried out to predict its fluctuation and forecasts the future price. Can daily news be one of them?

Daily news can sometimes have dire effects on everyday life. Whether they be responsible for the rise or fall of a stock value. Natural language processing consists of several state-of-the-art algorithms that has the capability to extract meaning out of texts. We have used majority of the NLP algorithms, from some primitive techniques to the currently used state of the art algorithm. On top of that, we have used a baseline classification algorithm to predict the fluctuation of Dow Jones Industrial Average (DJIA) index value - 0/1.

"1" when DJIA Adj Close value rose or stayed as the same;

"0" when DJIA Adj Close value decreased.

Additionally, we have also used the sentiment of the news to improve the predictions. They were hence broken down into Positive, Neutral and Negative groups. Working of this technique can drastically improve the stock predicting strategy.

II. RELATED WORK

Several approaches have already been tried out to predict stock fluctuations using text data for instance - sentiment analysis of social media data (tweets and trends), news, finance magazines headlines etc.

Most of these have successfully been able to predict the fluctuations up to some extent. Our work will add another perspective and technique of using news to predict the future trend.

III. APPROACH

Initially, we have performed sentiment analysis on each of the 25 headlines for every data point. They were counted and broken down into 3 additional features - Positive, Neutral and Negative.

All the news headlines for a particular day were then joined to form an overall News column.

1. Preprocessing

HTML and punctuations were rectified first. Stop words were then removed. Ideal preprocessing techniques such as Stemming and Lemmatization were then used to get an overall Cleaned News.

2. Vectorization Techniques

Bag of Words - It's a representation of text that describes the occurrence of words within a document. It involves two things: a vocabulary of known words and a measure of the presence of known words. It is called a "bag" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

TF-IDF (Term frequency - inverse document frequency) - It's a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Term frequency -

$$tf(w_i, s_i) = \frac{\text{no. of times } w_i \text{ occurs in } s_i}{\text{total no. of words in } s_i} \quad (I)$$

idf(w_i, D_c) =

$$\log(\text{no. of docs} / \text{no. of doc containing } w_i) \quad (II)$$

where w_i , s_i and D_c are word i , sentence i and Document c respectively.

Word2Vec - is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

Bert Transformers - Its key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The official paper's results show that a language model which is bidirectional trained can have a deeper sense of language context and flow than single-direction language models.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task.

3. Classification techniques

As this was an experiment to use daily news text to predict stock fluctuations, we have used a single baseline model for every vectorization algorithm which is Logistic Regression. Along with that, we have also experimented and included simple ANN on BERT vectors and tried to improve the results further.

4. Evaluation

Evaluation was done on the Dow Jones Industrial Average (DJIA), Dow Jones, or simply the Dow, is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the United States. Two labels were formed based on its fluctuation value -

"1" when DJIA Adj Close value rose or stayed as the same;

"0" when DJIA Adj Close value decreased.

5. Corpus

News data was crawled from Reddit World News Channel. They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)

Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01)

Dataset consists of 27 columns. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25".

6. Performance Metrics

Accuracy was used as a performance metric. This was done to get the number of correctly classified data over all the data. ROC-AUC score would have been a nice option too. Further evaluation will include that also.

IV. RESULTS AND DISCUSSION

Bag of words

Using Logistic Regression as a classification technique -
For $C = 0.001$ and penalty = 12, Validation accuracy we got was 0.5510835913312694.

Test accuracy we got was 0.5079365079365079.

TF-IDF

Using Logistic Regression as a classification technique -
For $C = 0.1$ and penalty = 12, Validation accuracy we got was 0.5541795665634675

Test accuracy we got was 0.4973544973544973.

Word2Vec

Using Logistic Regression as a classification technique -
For $C = 1$ and penalty = 12, Validation accuracy we got was 0.5572755417956656.

Test accuracy we got was 0.5052910052910053

Bert Transformer

Using Logistic Regression as a classification technique
For $C = 0.1$ and penalty = 11, Accuracy we got was 0.47883597883597884.

Using ANN as a classification technique. After 5 epochs -
Validation loss: 0.6828 and Validation accuracy: 0.5820
Test accuracy we got was 0.5026455026455027

Discussion

The results are decent but not much reliable. Fine tuning models further or using more advanced classification techniques or running the ANN model for more epochs could have improved the results further but the chances are low.

Given how advanced Bert is compared to other primitive techniques, the results are not up to the mark.

V. CONCLUSIONS AND FUTURE WORK

Given how unpredictable stock market can be, using news as to determine its future fluctuations can be helpful up to some extent but not completely reliable. This approach will surely assist the share market holders and traders to determine the stock fluctuations for a particular day. The availability of more news including financial news along with some other social media texts would have further confirmed the authenticity of this approach. This approach, for now, is very basic but has a great potential.

Further work consists of –

- Using more data. e.g. Financial news, tweets etc
- More fine tuning can be done.

- More advanced classification techniques can be included.
- Combining news data along with daily stock shares values - opening price, low value, high value, volume of shares and even the fluctuation of previous day closing price.

REFERENCES

- [1]. Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [May 23 2020] from <https://www.kaggle.com/aaron7sun/stocknews>.
- [2]. Base Bert Model with Tensorflow Hub by V.prasanna Kumar <https://www.kaggle.com/vpkprasanna/base-bert-model-with-tesnsorflow-hub>.
- [3]. Devitt, Ann & Ahmad, Khurshid. (2007). Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. Proceedings of the Association for Computational Linguistics (ACL).
- [4]. BERT Explained: State of the art language model for NLP by Rani Horev. <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>.
- [5]. A Gentle Introduction to the Bag-of-Words Model by Jason Brownlee.
- [6]. Dow Jones Industrial Average.
- [7]. https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average
- [8]. Applied Machine Learning Course <https://www.appliedaicourse.com/>.