Stock Market Trend Analysis Based on News

Mengqi Wu
New York University
Brooklyn, NY
mw4259@nyu.edu

Yuan Li
New York University
Brooklyn, NY
yl6606@nyu.edu

Dongzi Qu
New York University
Brooklyn, NY
dq394@nyu.edu

1. Introduction

Predicting stock market is a fundamental and important problem in natural language processing (NLP). There are several traditional methods for stock price predicting, which focus on historical prices. Recently, Bidirectional Encoder Representations from Transformers (BERT)[1] and its related versions have received wide attention from scientists. They can capture semantic and syntactic information in a sequence. As a result, sentiment analysis using BERT and data from social media and companies, such as Yahoo Finance and form 10-k, have received wide attention in industry. So we want to research if there is a better method to combine stock price data and news data to train a model that may help investors do a better decision.

2. Dataset

In order to implement one model predicting the trend of stock market benefited from NLP techniques, we intend to use textual data: form 10-k and daily financial news from Kaggle. Meanwhile, general financial headlines from Kaggle are also considered as an auxiliary to represent the trend of the whole market.

Besides textual data, historical price data represents the change of price in a continuous period of time. We use stock market dataset created by Boris Marjanovic. It will be helpful for us to generate labels for the company's stock price change.

3. Method

We will use NLP technology on both text data from 10-k forms and social media.

At first, we need to preprocess and clean up the raw data from text dataset: remove the html tags and all tables with numeric values from the dataset and make all text lowercase. Besides, we will also need to lemmatize all the data and use stop words to exclude some common words from further analysis.

For text files, we need to transform textual values into numeric values. We can generate sentiment bags of words from the dataset using the Loughran-McDonald sentiment word lists which is specifically built for financial textual analysis and generate sentiment term frequency - inverse document frequency (TF-IDF) to assign a score for each word. Then, we can get vectors to represent the textual dataset.

Finally, we can apply the sentiment analysis model on the dataset to predict the trend for stock market. The current sentiment model we consider use is Transformers[3]. It provides thousands of pretrained models to perform tasks on texts and can be modified easily.

4. Evaluation

The label we want to use to see how our model works is the company's stock price change. The evaluation method is modified based on Lee et al. (2014)[2]. We define three types of labels, which as the UP/DOWN/STAY label. We subtract the open price on the *1*-th day from the *5*-th day, to calculate the price change. 1.5% increase or more will be set as UP, 1.5% decrease as DOWN and between -1.5 % and 1.5% as STAY. Then the Accuracy and Micro F1 will be used as the evaluation criteria.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky. On the importance of text analysis for stock price prediction. In *LREC*, volume 2014, pages 1170–1175, 2014.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.