Sentiment Analysis for Financial Market Signals

Group members: Yuejiao Qiu, Ruoting Shen, Yuening Wang, Xiana Zhang

Problem Description

Market sentiment is one of the most important driving factors in the financial market, since it reveals how investors view the market, and stock prices will change so as to reflect these opinions on how the market will evolve. Nowadays, with the ease of information transmission, social media platforms contribute to the most financial market sentiments, as it allows people to share and express their ideas freely. Twitter has always been a platform where individuals are allowed to share information, such as news, politics, education or even financial guidance. There were multiple occasions, influential fellows were able to manipulate the financial market by their tweets.

Identifying market sentiments using text data on social media platforms have gained popularity, especially after a price surge on GME (GameStop) stock price in January 2021. Before the price rise, investors on Reddit reorganized that many institutional investors were betting against GameStop by shorting its stocks, which finally resulted in a large number of Reddit users purchasing GME stocks massively, making it a fourteen-fold price increase. This has suggested that market sentiment is of crucial importance to stock price fluctuations.

Therefore, NLP is a common choice that people opt to deploy to explore financial text data and sentiments. Previous research has been conducted in this domain, but mostly focusing on the formal texts, such as news articles and reports. However, posts and comments on social media platforms are full of internet slangs, jargons, and abbreviations, which makes it difficult to use transfer learning to analyze the discussions online that are relatively informal. In this project, we therefore aim to perform sentiment classifications on the financial tweets, trying out different methodologies and architectures, such as Bag-of-Words (BOW) model and fine-tuned BERT.

Methodology

Dataset

The data used for this project are Financial Sentiment Dataset (https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis) and Sentiment140 (http://help.sentiment140.com/home). The Financial Sentiment Dataset contains 5,322 tweets with sentiment labels (whether the tweet is “positive,” “negative,” or “neutral”). Moreover, there is also a significantly larger dataset, which is expected to be used for pretraining the BERT model. Sentiment140 contains 1.6 million tweets, as well as the information about the polarity (0: negative, 2: neutral, 4: positive), id, date, query, user, and text of the tweet.

Modeling Approach

The modeling approaches intended to use in this project are Bag-of-words and BERT.

Bag-of-words (BOW)

The first model we consider is BOW, which is relatively straightforward and traditionally prevalent in NLP. It will vectorize the tweets into a multiset of words, and each word exists independently. After vectorizations, the vectorized text will be fed into the designed neural network. But elements such as grammar and order are often ignored in the process. Therefore, we also consider some more advanced models.

BERT with the Financial Sentiments Dataset

One of the major models is the conventional BERT model directly trained and tested with the Financial Sentiments Dataset. Additionally to the basic BERT model, the parameters within the BERT architecture are unfreezed, and they will also get updated within each training epoch. And drop out layers will also be added to prevent the model from overfitting.

BERT transfer learning

A major obstacle of the previously mentioned model is the limited size of data, as the Financial Sentiments Dataset only includes over 5,000 tweets. To improve, we also proposed to adopt transfer learning of the BERT model. The BERT model will be firstly trained with fine tuning on the Sentiment140 dataset. Then we will record the BERT parameters obtained, transfer these parameters and use them in the BERT model for Financial Sentiment Dataset.

Evaluation

For general model comparison, loss, accuracy, and f1-score will be computed on the trained model on both train and test sets to evaluate the system. The BOW model will be included in the evaluation, so as to determine the best model.

Besides, given the context of this project which is based on the financial market, the association between the stock prices and sentiments will also be considered as one of the evaluation criteria. This can be done either with a general market index, such as S&P 500 index or Dow Jones Index, or focus on a specific stock.

Further Considerations

A key concern regarding our experiment is the availability of financial tweets. As mentioned previously, the Financial Sentiment Dataset contains only 5,322 tweets, which might not be sufficient to fine tune the BERT model. However, we also propose several solutions to address this issue, one of which is to first train our BERT model with fine tuning on a larger dataset, Sentiment140 and then transfer these parameters and use them in the BERT model for Financial Sentiment Dataset.

For the future plan, after fine tuning the BERT model and conducting the sentiment classifications, if given the chance, we would also be happy to continue our project to predict the stock market and come up with the strategies.