EE5183 FinTech Final Project Report- (Group 1)

Hao-Chun Chang(r08922125), Chih-Wei Ku(r08942027), Hsin-Yuan Hsieh(r08945021), Chih-Wei Chen(r08921076), Po-Ju Wu(r09921046)

Abstract—Recently, the web has rapidly emerged as a great source of financial information ranging from news articles to personal opinions. Data mining and analysis of such financial information can aid stock market predictions. We proposed LSTM and GRU model to predict stock closing price because both of them were powerful with sequence prediction problems. There were two parts of our input data, sentiment analysis and technical factors. The main social data was crawled on PTT, the largest forum used by Taiwanese, we gathered the data of TSMC from 2018 to 2020, which contains news and technical analysis factors. There are several technical analysis factors that we use, such as moving average (MA), KD, RSI, MACD, standard deviation of stocks. We found that sentiment analysis is indeed positive for predicting the closing price, although there is a lot of noise in stereotypes from PTT, we still confirmed that the sentiment of PTT can indeed predict the Taiwan stock market quote change to some extent. Furthermore, after experiments, GRU model had been shown to exhibit better performance on certain smaller and less frequent dataset than LSTM model.

***Index Terms*—Fintech, Stock prediction, Sentiment analysis, LSTM, GRU.**

I. Introduction

As web-based technologies continue to be embraced by the financial sector, abundance of financial information is becoming available for the investors. Mining social media data to forecast the future has been a popular research in recent years. With the development of social media, large amount of information can be derived from the social media. Taking PTT for example, it's a popular BBS in Taiwan and there are over fifteen thousand users sharing their opinions at peak time, the information on social media not only tell us what is happening in the world bot also the author's feelings toward an event.

Stock market prediction has attracted the attention from academia in recent years. The trend of stock market has been considered to be related to the emotion of crowd. Hence, how to analyze the feelings of crowd is the issue worth discussion. More and more people are willing to share their experience and opinions of happenings on social media, similarly, investors also express their feelings on social media. Forecasting stock market behavior through mining news or social media information has become a popular research topic.

The main goal of this study is to forecast the stock market behavior in Taiwan and discuss the correlation between online news emotion and stock market.

II. Material and Method

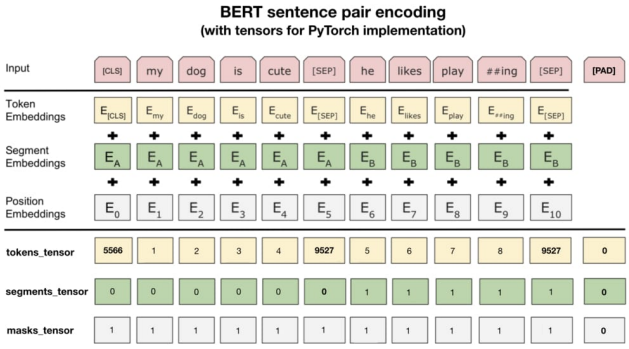
The predict stock target that we choose is TSMC, one of the famous company about supplying semiconductor around the world. There were be two parts of our data: 'sentiment analysis' and 'technical index'.

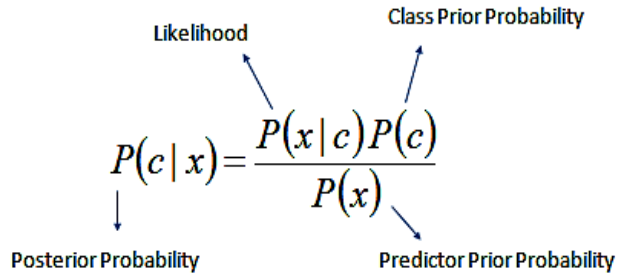
A. Sentiment analysis:

We crawl the news topic that mention TSMC on PTT in 2018-2020, then deal news topic with natural language preprocessing by using the 'BERT' and 'SnowNLP' model to figure out the sentiment score from topic.

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. In order that BERT is only suitable for English, we translate the Chinese news topic data into English by google translator api, although the topic includes some technical terms in financial stock market, it still make a precise translation.

SnowNLP is an useful library for traditional Chinese natural language processing, it generates the sentiments score from each article's topic by native bayes classifier. Below are the concept of these two natural language processing tools.





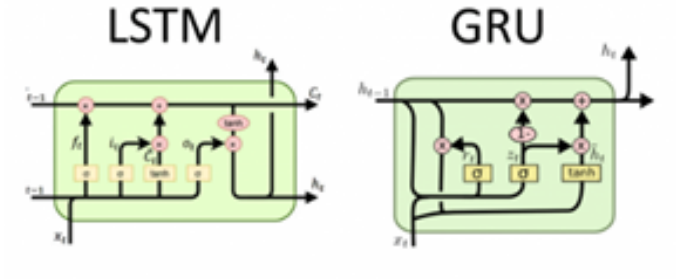
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 1. Snownlp is base on native bayes method to classify the sentence into sentiments score between 0-1.

B. Technical Index

We gathered data of TSMC from 2018 to 2020, which contains news and technical analysis factors. There are several technical analysis factors that we use, such as moving average(MA), open, close, high, low, KD, RSI, MACD, standard deviation of stocks. The data is normalized for training.

We propose to apply an LSTM model because it is powerful with sequence prediction problems, and the LSTM model could store past information. There will be total 14 features in our modeling input. Furthermore, we determine to apply GRU, a LSTM with a forget gate, but has fewer parameters than LSTM. GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets.



III. EVALUATION

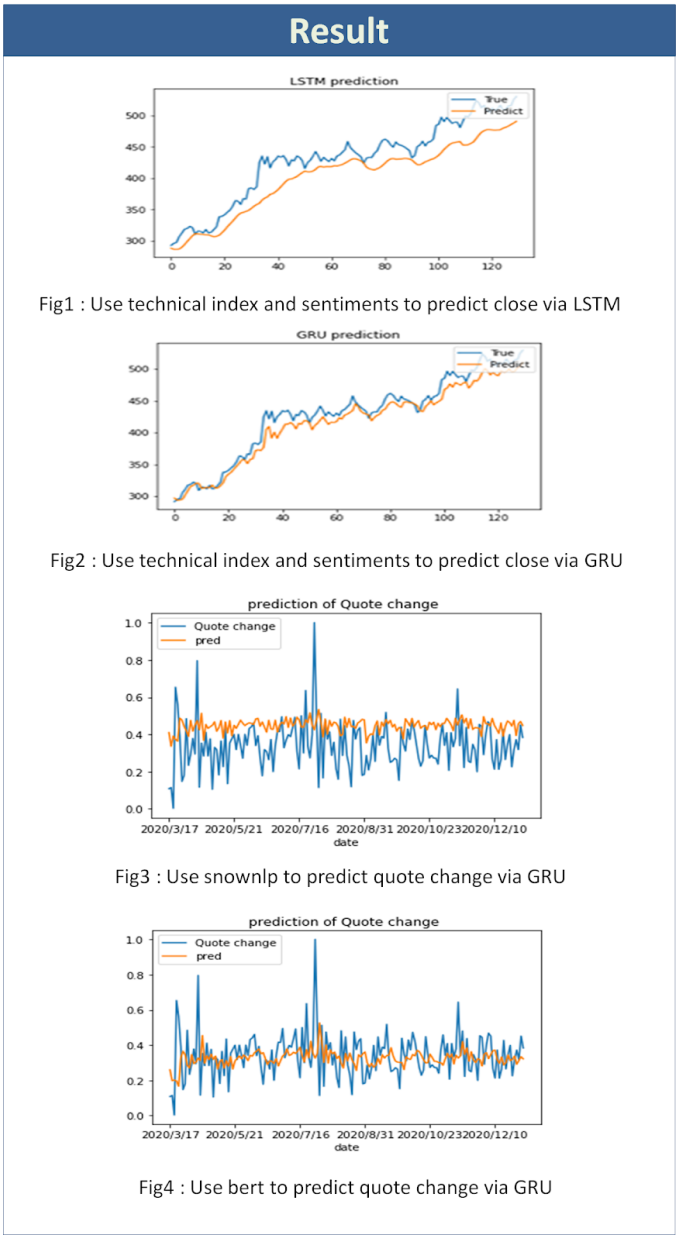


Fig1 : Use technical index and sentiments to predict close via LSTM

Fig2 : Use technical index and sentiments to predict close via GRU

Fig3 : Use snownlp to predict quote change via GRU

Fig4 : Use bert to predict quote change via GRU

In our experiments result, we find that the input features including 'BERT' and 'SnowNLP' can make the predict closing price closer to the true closing price. And GRU model perform better than LSTM in this case, but both of them cannot predict the true closing price perfectly. As the result, we hope to obtain the correlation between stock quote change ratio and sentiments score. Figure 3 and Figure 4 shows that BERT has higher correlation with closing price than SnowNLP, we consider that the main reason is because the default pre-training dataset in SnowNLP is about buying/selling comments, there must lost some financial words and terms in the dataset. However, BERT is pretrained by the social posts on Twitter, which contains lots of categories of word, there may involve financial terms in the pre-train dataset.

IV. Discussion

2020 is a tough year for every investor and stock market, if we hope to get more precise performance, there are some improvement we can discuss in the future.

First, we only obtain 465 days PPT information in dataset, and split 300 days for training set, the others for testing set. The small dataset may cause training difficult, we shall crawl more data to revise the result. Second, 14 features are not enough for predicting stock price obviously, a better way for natural language processing is to use word-vector method to do analysis, using more vectors and features to make prediction may have different result. The last improvement is to train SnowNLP dataset by our stock news, in order to make the sentiments score has more correlation with the true stock price effect.