Team ikun:
Yixing Zheng (yixingz3@illinois.edu) - team captain
Zerui Tian (zeruit2@illinois.edu)

Text Mining on Financial Text for Predicting Stock Prices

I. Project details

Our topic is about “Text Mining on Financial Text for Predicting Stock Prices”. We hope to apply text mining technology to the financial field, perform data mining on some specific financial data texts and use the results to predict stock prices. We want to determine the correlation between stock price fluctuations and StockTwits, and we want to compare different data processing methods and models to get the best model to make predictions.

This is interesting to us for two main reasons. One reason is that there are not many applications of text mining to financial markets, and many of them are related to sentiment analysis, and do not directly link text data and stock price data. So this is actually a new attempt for us, and we also think this topic is very challenging. Another reason is that we would like to explore the utilization of machine learning/deep learning algorithms in natural language processing and extend, or at least validate, that machine learning/deep learning is able to improve the performance for some NLP tasks.

Approach wise, we will first use some methods to label the financial article data. These labels are based on changes in stock prices, which can be used as 1 for rising and 0 for falling. We will then try different labeling methods. And we will try different machine learning models and deep learning models, such as Naive Bayes or support vector machines based on the TF-IDF model, and deep learning methods based on word embedding. If time permits, we will also try transformer-related models such as BERT.

Our goal is to find a proper labeling method, and a model that performs the best among many machine learning and deep learning models. These models can better correlate textual data with stock ups and downs and make more accurate predictions.

For the evaluation, since we have performance metrics - accuracy, F1 Score and so on, we will be able to directly compare them and evaluate our work.

II. Tech stack and dataset

We plan to use Python and some machine learning libraries (numpy, pandas, pytorch, etc.) in Google Colab. The dataset is text data from Stocktwits, and will be updated to our GitHub repository.

III. Planned tasks and workloads

- Finding and processing the dataset - 2 hours
- Initial research, brainstorming and discussing which algorithm to use - 2 hours per person
- Dividing the task and setting up the environment - 2 hours per person
- Working on the algorithm, including training initial smaller dataset, evaluating model with larger dataset, training with larger dataset then evaluating the model with larger dataset, and documenting the code. We will try multiple machine learning and deep learning models and compare the results of different models. - 17 hours per person
- Administrative tasks, including setting up CMT and GitHub for submission, drafting the project proposal, scheduling meetings for progress updates, drafting progress reports. - 4 hours for team captain
- Preparing the demo, including the slides and presentation. - 1 hour per person

All the above sums to a total estimate of **50** hours for two team members with potentials of encountering issues, such as environment setup, incomplete or invalid data fixing, and initial approach doesn’t perform as expected and requires pivoting, which might add another 5 to 10 hours of work for the team.