# Using News and Historical Trend to Predict Chinese Stock Movements

Kunsheng lyu, lvksh@umich.edu, 41190256

Ivy Wang, iwango@umich.edu, 50978422

Zeyuan hu, zeyuanhu@umich.edu, 09972157

Predicting stock movements has been an exciting topic for a really long time. Before natural language can be formally studied, predicting future stock movements based on the historical trend of the stock price time series is the mainstream method. However, it's almost impossible to predict some drastic change based on history (otherwise it won't be a 'drastic change'). In recent years, with the advent of natural language processing, researchers have started to look at the text data behind the stock markets and try to make better predictions using new techniques. Indeed, there are so many stock managers who studied the company news and financial reports to help them have a better insight towards the stock. An interesting question is, can we imitate those stock managers and use machine learning techniques to help us "read" all these complex news and reports and tell us the future trend of certain stocks? We are excited to have a try!

As for why we choose Chinese stock market instead of the US stock market, we notice that in the US stock market, "Individual investors made up just 10% of the market's trades in 2019", quote from *Business Insider*. But there are more retails in Chinese stock market, about 99% (Really surprising!), which means that the news is more likely to affect those people and have the ability to cause interesting changes in the stock market.

The main objective we trying to predict is 10 days return, which is defined as

$$R_t = \frac{x_t - x_{t-10}}{x_{t-10}}$$

, where $x_t$ is the close price of time t, and $x_{t-10}$ is the close price of time t-10. We will construct our data frame in a rolling manner to extrapolate our sample size. The measurement metric we use is MSE, which is defined as

$$MSE = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

, where $y_i$ is the ground truth 10-day return series, $f(x_i)$ is the predicted 10-day return series.

To get the data we need, we plan to crawl historical stock prices as well as historical news contents from a Chinese finance website https://finance.sina.com.cn/ . It contains all the information we need and it's formatted as a static website so it's easy to crawl. We decided to use the data generated from the past 10 days to predict the future 10-day return, so our dataset will contain the information from the past 10 days, including text and finance factors, in each row (Each row represents a certain stock on a certain date, with the label being 10-day return starting from this date.)

For this project, we divide it into several steps:

1 Data Retrieval: Using crawling methods to crawl all the data we need for this project.

2 Data Cleansing: Data from the crawling process may be messy and hard to use, so we need to preprocess the data first.

3 Numerical Feature Engineering: It's hard to predict the stock using solely the text, so we need to construct some stock factors like 5-day variance, etc.

4 Text Feature Engineering: We plan to utilize Deep Learning technique to preprocess the text, so we need to find a suitable network framework and implement it on Pytorch.

5 Model Building: We will try different machine learning regressors and compare their performance.

6 Model Applications: If we have time, we will extract some up-to-date data and validate our model on real data.

The steps are listed above, for crawling the text data, we will use request and beautiful soup in python, and the further NLP and machine learning process, we will use PyTorch and NLTK, gensim,sklearn,lightgbm or any needed packages and frameworks.

In this process, we three will work together to have the best learning experience out of this project.

We listed some tasks and steps above. All the steps are related close to others, so it will be potentially hard to make a clear distribution of the tasks to each person. We might take two tasks per person and work together closely with each other to make the best outcome.