**Supplemental Files**

1. **Data source for experiment**

Supplementary table 1-AAPL

Supplementary table 2-MSFT

Supplementary table 3-KO

The feature of the comment dataset is that it contains stock comments of three companies in short-term.

The column 1 is the date, from January 2021 to may 2022.

The column 2 is the titles of stock comments.

The column 3 is the stock suggestions given by authors of the stock review: buy, sell and hold.

The column 4 is the emotional value of stock evaluation. Its range is -1~1, 0: neutral -1:selling, and 1; buying.

The column 5 is the difference between the share price seven days after the release of the stock evaluation and the share price on the day of the release of the stock comments.

The column 6 is the share value on the day of the issuance of the stock evaluation.

The columns 7~12 are the daily forecast values of five models.

The column 13 is the number of trading days from January 1, 2021.

1. **Data collection from StockTwits**

API: <https://api.stocktwits.com/developers/docs>

Data dictionary: Fin\_word\_v1.0.zip

There are 7 parts for each token (word), and they are saved in json format.

“token”: word

“bull\_freq”: frequency in bullish set.

“bear\_freq”: frequency in bearish set.

“bull\_cfidf”: collection frequency in bullish set.

“bear\_cfidf”: collection frequency in bearish set.

“chi\_squared”: chi squared test result of the token

“market\_sentiment”: calculated by bullish PMI minus bearish

E.g. { 'token': 'buy', 'bull\_freq': 14489, 'bear\_freq': 1592, 'bull\_cfidf': 61.539806954702385, 'bear\_cfidf': 52.32250663139482, 'chi\_squared': 14711.705215251208, 'market\_sentiment': 0.5961743093876137, 'word\_vec': [0.0928284227848053, -0.10893399268388748, 0.12348346412181854, … , -0.01443735882639885] }