News Sentiment Analysis for Stock Data

Group 32: Changyu Wu, Hanyu Zhang, Felix Chung

1. Description

As one of the most important financial security, stock is a great option for financial investment. However, predicting the trends of stocks is not an easy task. Stock price is mainly determined by a group of buyers and sellers that construct the market, and many of these buyers and sellers are ordinary people who trade irrationally. Most ordinary people will not conduct full financial analysis of the company and trade security based on psychological and social factors. Here, we decide to focus on the most significant influence factor of stock price variation, news, to discover the relationship between news and stock price. In our project, we will analyze sentiment of news headlines, labeling them with positive, negative or neutral tags, which represent predictions for stock price changes: up, down, or staying the same. For future use, we can collect recent news into our dataset to make predictions as references for our investments.

2. Dataset

Our dataset is available on Kaggle. To collect more data for better analysis, we merged available .csv files all together. Here is the original link:
https://www.kaggle.com/sidarcidiacono/news-sentiment-analysis-for-stock-data-by-company

The dataset includes three columns and over 15k rows. The first column is the sentiment label, ranging between 0, 1 and 2. Zero means the stock went down by market close the day the article was published (negative). One means the stock went up by market close the day the article was published (positive). Two means the stock stayed the same by market close the day the article was published (neutral). The second column is a ticker symbol for the company. The third column is the news headline.

While we could not get date information on each news headline, we think it should not be a problem for sentimental analysis.

3. Methodology and Expected Result

To understand our dataset, we will use some EDA (Exploratory Data Analysis) methods, preprocess the raw data with numpy, pandas, nltk, re, etc., and draw plots to visualize it using matplotlib or flourish.

For sentiment analysis, we will first perform tasks like removing stop words, tokenization, feature extraction with TF-IDF (Term Frequency-Inverse Document Frequency) and other related methods.

We will be using three models: NB (Naïve Bayes), LSTM (Long Short-Term Memory) and BERT to predict the trend of different stocks with text. We will begin with NB (Naïve Bayes), a less complex model compared to neural networks, to be the

baseline of our final task. Then, we will implement deep learning models like LSTM (Long Short-Term Memory) and BERT to make a prediction and compare results from different models. To evaluate our model performance, we will use the F1-score.

For further study, if feasible, we would try to implement additional classification models such as SVM, also to use keras or other related tools to construct our LSTM model for prediction, instead of using python libraries directly.

4. Timeline

Week 9: EDA

Week 10: Data preprocessing, implement models

Week 11: Implement models, tune model

Week 12: Report Write up

Week 13: Presentation

5. Responsibilities

Since we only have three people, it would be possible to do most of the tasks together. To be specific, Changyu Wu will be primarily responsible for data preprocessing and implementing models. Hanyu Zhang will take charge in EDA and model tuning. Felix Chung will implement models and write reports.