Project name: BetterThanCNN
Team name: Fu's Fantastic Four
Members: Chenxuerui Li, Jianyao Fu, Senhao Hu, and Ziwei Jiang

Project Summary Sheet

Highlights

We put our emphasis on the Natural Language Processing with Machine Learning models. This process is incredibly tiring and time-consuming, but due to the limit of 5-minute (or 10-minutes to professor and TAs) presentation, we could hardly look into that thoroughly. Thus, we hope that you could focus on the **Stock-News Modeling part** of this summary sheet.

One thing to notice is that Our model prediction wasn't based on the sentiment analysis, but the word frequency feature.

Motivations & Data Description

When betting the market trend and implementing trading strategies, traders on the street always utilize the current market news as a main signal to detect the major market movements and volatilities. Given a relatively efficient market, the news tends to be absorbed and reacted by market players within several hours.

Thus, this project aims to predict the US stock market performance based on market news headlines. This project makes use of a dataset referenced from Kaggle (https://www.kaggle.com/aaron7sun/stocknews). The dataset is composed of top 25 daily market headlines ranked by Reddit votes and Dow Jones Industrial Average (DJIA) from 2008 to 2016.

Data Exploration & Sentiment Analysis

With basic data exploration in the beginning, we found frequent words and 'crisis' related words. We also did words dispersion analysis and LDA analysis.

Then we made sentiment analysis based on daily news headlines. It turns out that the negative sentiment values take more percentage than the positive ones. It proves our assumption that newspapers tend to report more negative news to sell better.

Stock-News Modeling

In the data processing part, we first divided the dataset into training dataset and testing dataset. We selected data before 2015 as our training dataset and the rest as testing dataset. The train and test proportion are nearly 8:2. Then we joined the daily top 25 headlines into one string from which we could compute daily term frequency.

In the modeling part, we used two methods to get the term frequency of daily headlines, that is CountVectorizer and TfidfVectorizer. These two methods will count the frequency of each word in the headlines and convert the headlines into a word frequency matrix. Besides, TfidfVectorizer will also compute the inverse documents frequency to decrease the term frequency if a word appears too much in other headlines. We also used four machine learning models to fit the label and the term frequency: Logistic

Regression, Naïve Bayes, Random Forest and Neural Network. By applying GridSearchCV, we got the best parameters of each model to get the highest accuracy.

The initial accuracies of all eight models were all about 50%, which means that our models were no better than randomly guessing the vibration of the stock price. Then we decided to increase our threshold to 0.7 with the belief that people are more likely to react to the bad news than to the good news and only when the news is good enough the stock price will increase. By doing so, we got higher accuracy for each model and our best model is Logistic Regression with default parameters and CountVectorizer with ngram_range= (2,3) (using two and three connected words as model features.). Its accuracy is 57.94%.

Finally, we calculated the confusion matrix of our best model which showed that our model could predict 73% true label 0 and only 43% true label 1. Therefore, our model could be used as a reference for investors' short-selling strategy. Moreover, we extracted all the meaningful words with the most negative coefficients in the model and they could be used as an indicator for the decrease of stock prices.

Takeaways & Objectively Existed Model Bias

We believe there is still some objective bias for further research.

The first obstacle is the definition of a piece of news being good or bad. We concluded that there would always be an ingrained counter reaction among market sectors, since a news can be good and bad at the same time among different market sectors. Secondly, there would be a lagging between the news and the market reaction. This may lead to an inaccuracy when we pinpoint the correlation on a daily basis. Lastly, because the investors are risk-averse, they tend to overreact on the negative news. The media has also sensed this inclination and is more willing to publish more negative news.