**Brief Report**

Lue XIONG, Xinhui Gu, Tianyi Peng, Qingyao Sun

**1. Evaluation Metric**

The stock price data we used is MSTF price from 2014/3/27 to 2017/9/12. We split the data into three sets: training set from 2014/3/27 to 2016/8/29 with 611 observations; validation set from 2016/8/30 to 2017/5/9 with 173 observations; test set from 2017/5/10 to 2017/9/12 with 86 observations.

We would compare our model performance on the test set with a benchmark model and a naive model as detailed in section 2 and section 3 respectively. We would also evaluate our model performance with a p-value style metric on the test set. Concretely, we would count the number of correct stock price movement predictions produced by our model, and then calculate the probability that we observe this number or more correct predictions assuming the number of correct predictions follows a binomial distribution with n=86 and p=0.5.

**2.Benchmark Model:**

The naïve model focused only on the words with features ‘Positive’ and ‘Negative’. Our benchmark mode is counting the difference between positive and negative words each day. If the difference is within 0.3\*SD(All data) we predict a neutral return of 0. If the difference is greater than 0.3\*SD(All data) we predict a positive return of 1 and if the difference is less than 0.3\*SD(All data) we predict a negative return of -1.

**Data:**

Time range: (Mar 27th, 2014 to Sep 12th, 2017)

In total 1239 days of market news data.

**Data Cleaning:**

We excluded weekend, holidays and days where there are no related news. After data cleaning, we are left with 852 days of data.

**Results:**

The histogram is the difference between the actual return (+1 if positive and -1 if negative) and the baseline model predicted return which are also + or – 1. Correct prediction was 31.18%



**3. Naive Model**

In our training set, there are 310 up moves out of 611 observations. Therefore, under this naive model, we would randomly predict a up move on each date with probability 310/611=50.73% on our test set and then we would compare our prediction with the actual labels of the test set. It turns out that our naive model made 45 correct predictions on test set, with probability of correct prediction being 52.32%.

**4. State-of-the-art model:**

**Model:**

We used a multiclass support vector machine method to classify the returns as three types: positive, neutral and negative, where if the return on a particular day is within 0.3 times the standard deviation of the historical returns during the period, we set the return to 0 because it’s not significant enough. Resulting in a vector of 1,0 and -1 for our y variable.

**Feature:**

The features are: Negative, Positive, Uncertainty, Litigious, Constraining, Superfluous, Interesting and Modal. Where each feature is the number of times that words of that category appeared on a particular day in news headlines.

**Data:**

Time range: (Mar 27th, 2014 to Sep 12th, 2017)

In total 1239 days of market news data.

**Data Cleaning:**

We excluded weekend, holidays and days where there is no related news. After data cleaning, we are left with 852 days of data.

**Result:**

This is a histogram of the difference between predicted and the actual returns, which is the difference between two vectors of -1, 0, and 1s. As we can see, majority of the mass is around 0. To be more precise, the proportion of correct predictions were 0.735.

