Project 3 Group 1: Qixu Cao, Xiaoman Dong, Yigao Li, Taoran Yu

Professor: Ami Gates

ANLY 501

December 9, 2017

**The impact of Social Media on Stock Prices**

***Abstract***  
 Twitter, as one of the most popular social media platforms, has been considered as an influential factor to the stock market by many professional publications recently. The purpose of this study is to perform a deeper analysis and research about stock and twitter data. The approach combines hypothesis testing with machine learning. So it could be used to evaluate the accuracy of Twitter users’ prediction (which will be indicated by the sentiment score in their tweets) of stock prices. With machine learning, we are able to separate users into correct prediction and incorrect prediction. From the two separated groups, the additional analysis on the influential factors in predicting stock price of twitter users could be revealed.

***Introduction***

Social media is about expressing feeling and showing life, while stock market is about making money. The interaction between these two components is creating some interesting sparkles. Who does not want to be famous and rich at the same time? Previous study has already shown that there is some correlation-causation between social media and stock price, here, this research will testify the previous study’s conclusion again and generate hypothesis testing from other aspects of social media and stock prices, such as if verified users are more likely to express their feelings about stocks; if follower number has correlation with correctly predicting stock price changes. After completing the hypothesis tests, the research will apply the conclusion from hypothesis test and only takes the influential factors of twitter user information. For example, if the hypothesis testing shows sentiment collection from verified users are more likely to predict stock prices, then the next step of this research will apply machine learning sentiment analysis with verified users and calculate the accuracy again.

***Sentiment Analysis: First Attempt***

The most important analysis in the entire project is to find whether people can predict stock price trend in the future based on history Tweets. Null hypothesis is that people can use Tweets from history and stock data to predict future stock prices, otherwise the alternative hypothesis is Tweets can not used for prediction. To study sentiment analysis, machine learning techniques and many classification methods are introduced, such as SVM and Naïve Bayes (NB). We merged two original datasets by stocks. The merged dataset has three variables (“%volume”: volume percent change; “d1return”: stock return in Day1; “score”: sentiment score) and one class (“label”: stock price trend in Day2 as “Positive” or “Negative”). By default, 75% of the dataset is training data. A summary of using machine learning methods follows in the table below:

|  |  |
| --- | --- |
| Machine Learning Methods | Classification Accuracy |
| Support Vector Machine | 0.52 |
| Naive Bayes | 0.48 |

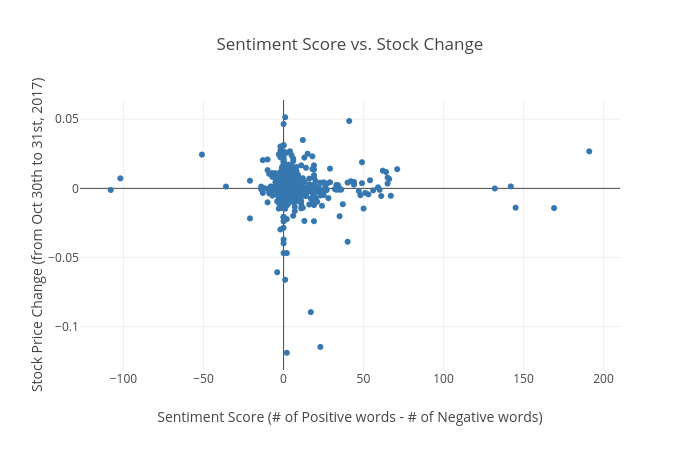
**Figure 1. Supervised Learning Analysis Results**

In general, the result is not gratified because all accurate scores are around 0.5, which means training machine with 75% of the data can only successfully predict future stock price trend with approximately 50% of chance. These machine learning methods are not even better than flipping a coin.

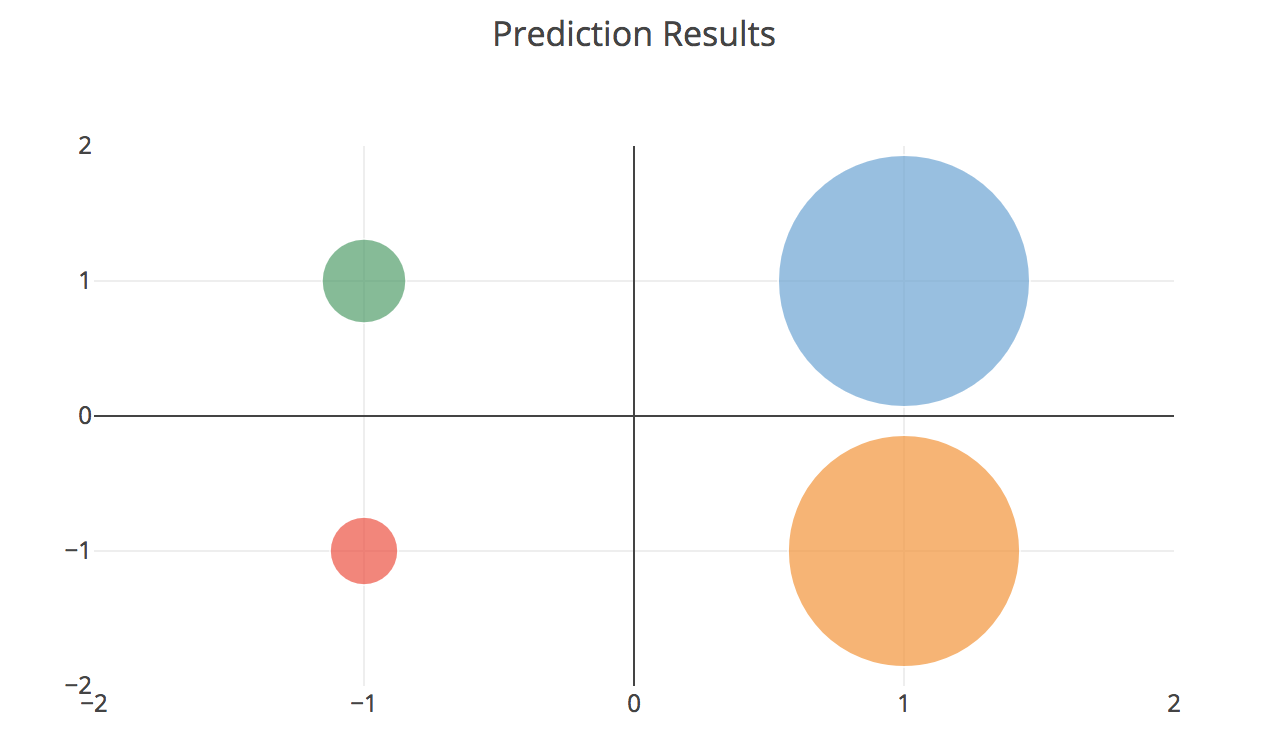
***Additional Analysis: Features of accurate stock prediction***

As mentioned above, we used several machine learning methods to find the correlation between sentiment scores and stock prices, however we failed to get good results. This time, our group wanted to figure out the features of an accurate prediction of stock based on the sentiment score and stock price changes. The process we used is demonstrated in our analysis below:

1. Collecting data: column “score” from new\_final\_twitter data.csv and “d2 return” from final\_stock\_data.csv. Here, score refers to the sentiment score of a tweet, which is calculated by the positive word counts minus the negative word counts. “d2 return” stands for the daily percentage change of a stock price.
2. Classifying data: In the “score” column, we viewed all the positive scores as 1, the negative scores as 0, and ignored data with score of 0. The “d2 return” column, shows the percentage of a stock price change. If the data in the “d2 return” column is greater than 0, we then classify this data as 1, while simultaneously classifying all data with a negative percentage as 0.
3. The reason we didn’t take into account scores equal to 0 or with a percentage of 0 was because this data couldn’t help us make predictions on stock. Our goal was to predict future prices. However, no percentage of 0 or score of 0 can predict or be used for prediction.
4. Analyzing data: We used all classified data which had (positive score, positive d2 change) and (negative score, negative d2 change) as our accurate prediction data, and classified all data with (positive score, negative d2 change) and (negative score, positive d2 change) as inaccurate prediction data. Now, we visualized the data by using a scatterplot to do the following:

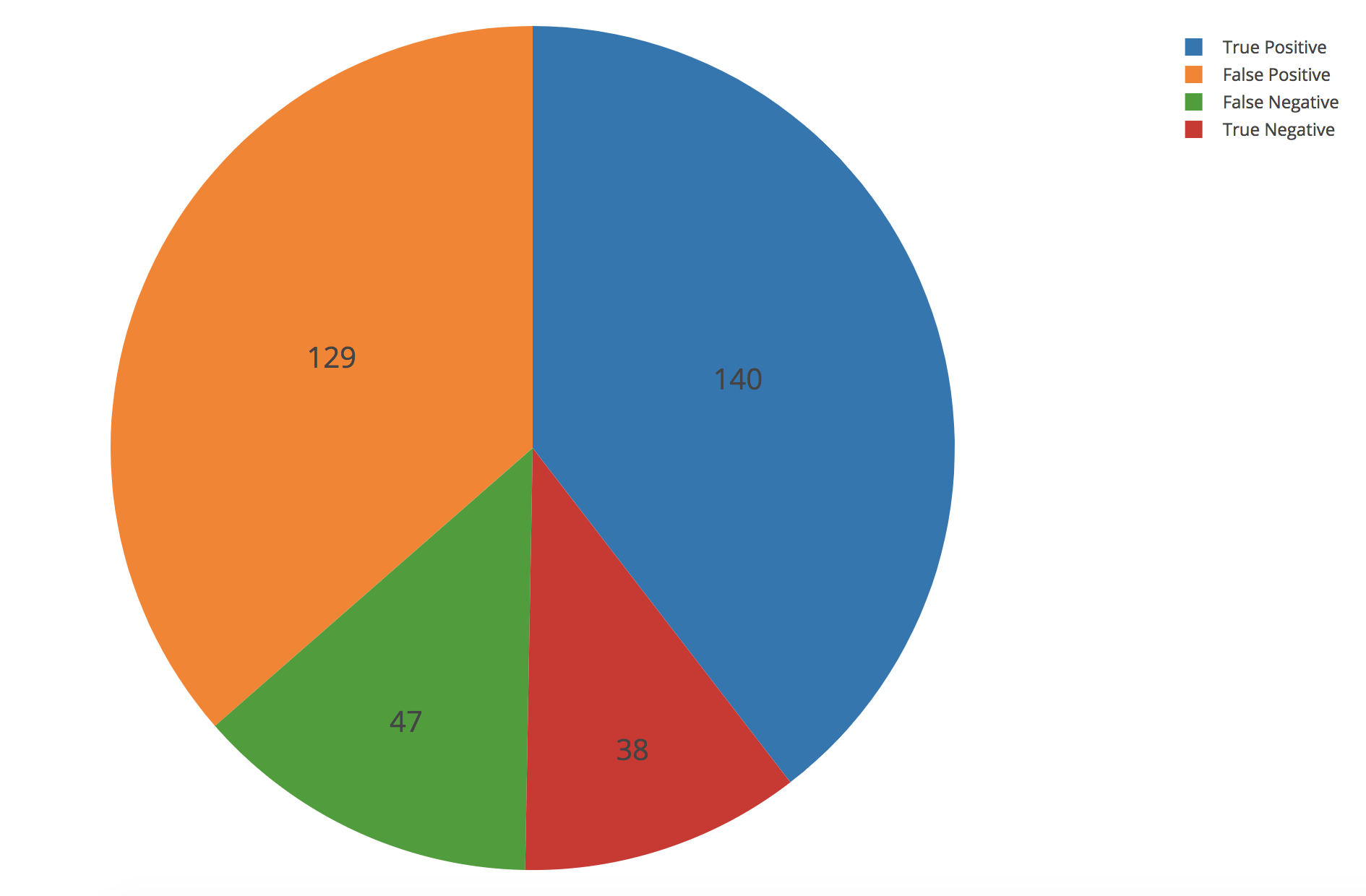


**Figure 2. Sentiment and Stock Return Relation**



**Figure 3. Visualizing Sentiment Analysis Results**

In the visualization Figure 3 and 4, we can see that there are 140 data with a positive score and positive percentage change, 38 data with a negative score and negative percentage change, therefore the total number of accurate prediction is 178. The number of positive scores and negative percentage changes is 129, and 47 values with negative scores and positive percentage changes. The total inaccurate data is 176. If we use the pie chart to represent the distribution, we can also see that the left side and right side of the chart is relatively similar, and weight approximately the same.



**Figure 4. Summary of Prediction Results**

|  |  |
| --- | --- |
| ***Accurate prediction data(Group 1)*** | 178 |
| ***Inaccurate prediction data(Group 2)*** | 176 |

**Figure 5. Separate Prediction Results to Groups**

We found out that the number of our accurate and inaccurate predictions were relatively close, which caused our group to apply machine learning methods like: SVM, Decision Tree, Nearest Neighbors, Naive Bayes, and Random Forest Classifier fail to get a good prediction. However, this time our group’s aims was to analyze the features of two groups. The results of this new analysis showed in the next section.

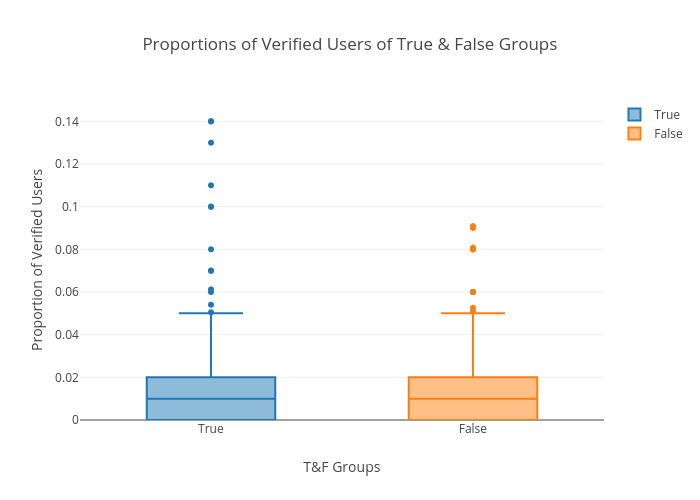
***Problems with Twitter data***

Twitter data contains tweets from various users who may be verified by Twitter. Twitter users have followers and those followers can view, like and comment on tweets. Our Twitter data may have meaningless or misunderstanding tweets that influence final predictions. Therefore, we constructed two new hypotheses:

1, Verified users can be a factor that impact on sentiment score

2, Number of Twitter followers can be a factor that impact on sentiment score

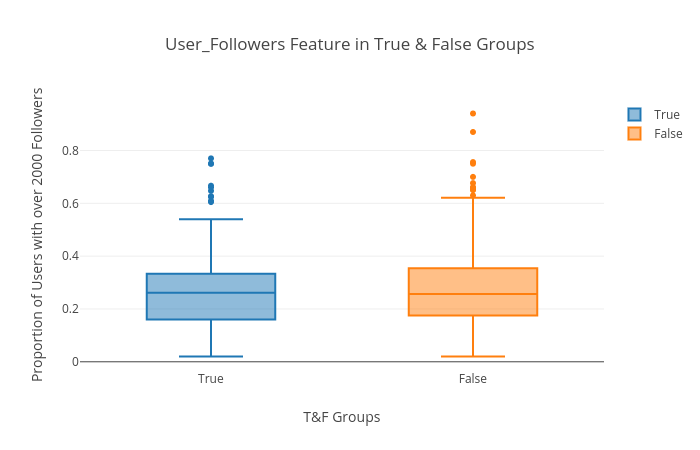
The first hypothesis is verified users tend to impact on sentiment score. We want to discover whether the proportion of tweets by verified users differs between Group 1 and Group 2. Therefore we used plotly to show the proportion distribution.



**Figure 6. Proportion Distribution by “Followers” Factor**

It is clear that the boxplot of true and false group have the same box, which only stands for a small proportion of users. Hence we focus on the scatter points above the box, specifically, only compare the points over 6% for both groups. From the plot, we conclude that the proportion of verified users is greater than the non-verified users. After constructing two-way T test, the test result of 0.0445, which is statistically significant. It tells that more tweets collected for Group 1 were posted by verified users than Group 2.

The second hypothesis is that Twitter user with larger amount of followers can predict the stock trend more accurate. By using plotly, we get the result as below:



**Figure 7. Proportion Distribution by “Verified User” Factor**

Here the true group refers to the group of users with accurate stock trend prediction, while the false group is the group of users with non-accurate stock trend prediction. From the graph above, we can see that the number of non-accurate prediction is greater than the accurate prediction. By doing T-test, we get the result as -1.09, and p-value as 0.276 which is statistically insignificant. Therefore, we can say that the number of followers is not a factor that impact on sentiment score.

***Results of Additional Sentiment Analysis***

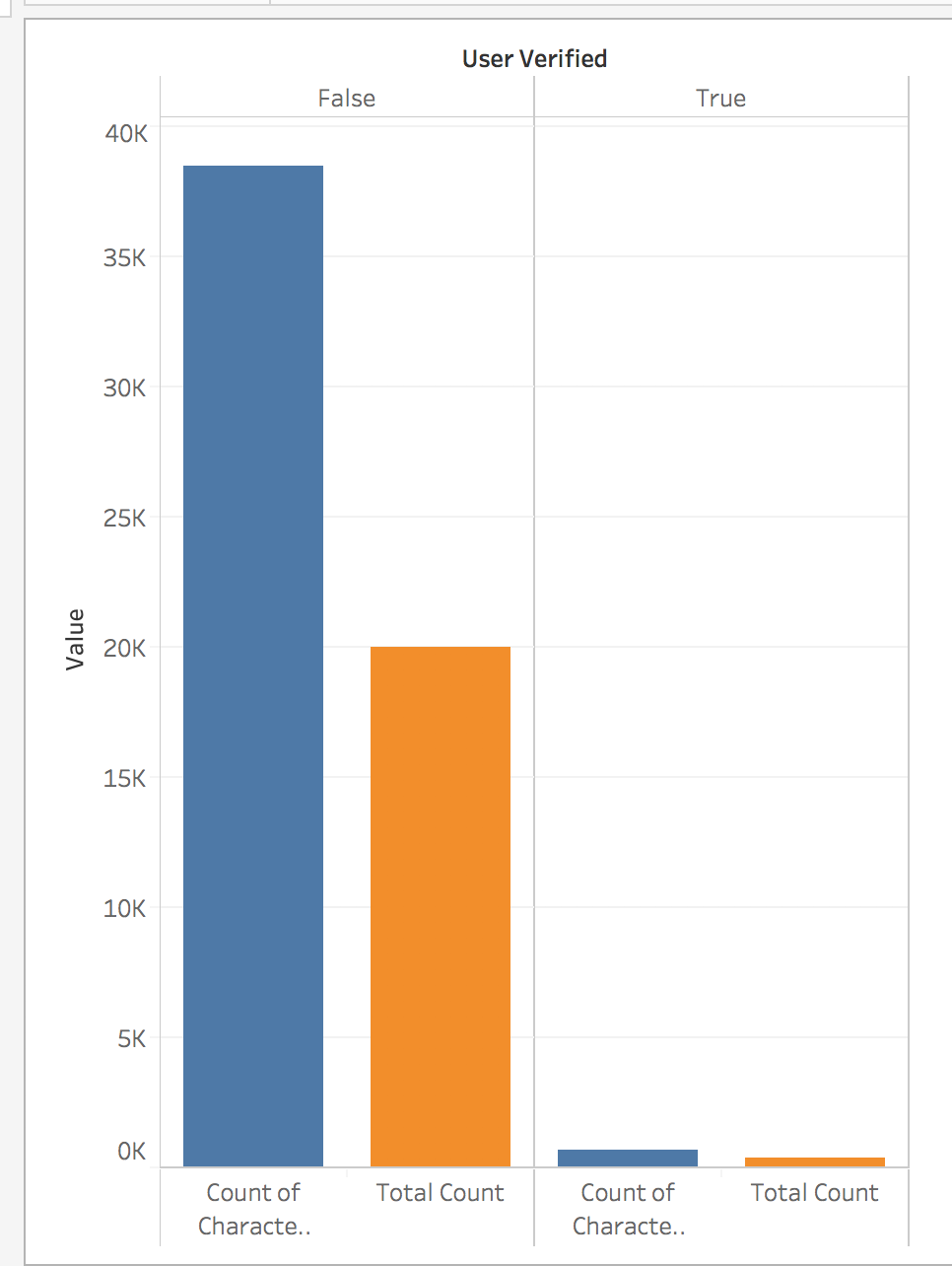
From project 2, we added two columns “total count” and “score”. Total counts stand for the total number of positive words and negative words in a tweet, and score means the number of positive words minus negative words since this shows how much a tweet shows its positive or negative attitude towards a company.

From the analysis we did in the previous section, we know that the verified user can be a factor which has an impact on stock price trend prediction by hypothesis test. Our group first collected new dataset from our API sources. After implementing the same way of merging data like what we did in the previous sentiment analysis, we applied the same analysis again. But this time, we filtered Twitter data to make sure that we have at least 6% of tweets which were posted by verified users. We got the new classification accuracy with naive bayes outcome of 0.66 and support vector machine outcome of 0.67.

|  |  |  |
| --- | --- | --- |
| Machine Learning methods | Old Classification Accuracy | New Classification Accuracy |
| SVM | 0.52 | 0.67 |
| Naive Bayes | 0.48 | 0.66 |

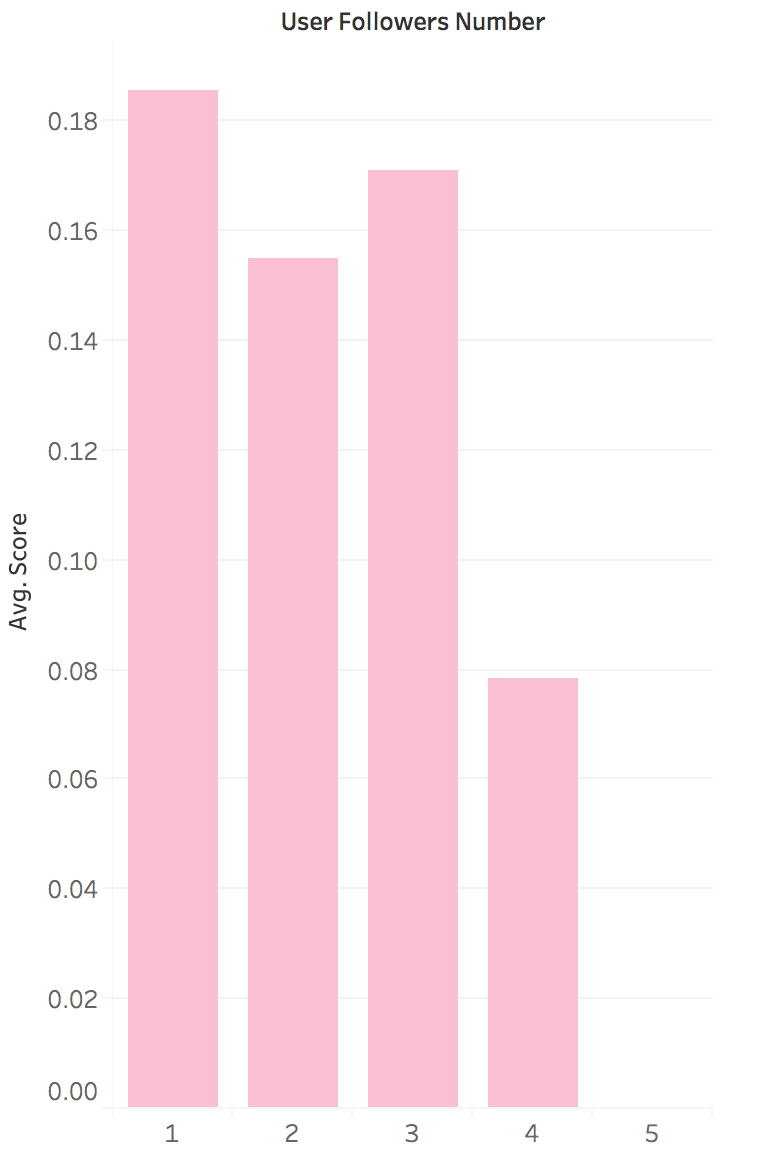
**Figure 8. Supervised Learning Results Comparison (Before and After)**

***Other Data Visualizations using Tableau***



**Figure 9. Distribution of Users by Verification**

The graph above shows the distribution of total count and count of characters among verified and non-verified users. As can be seen, the amount of characters and total count of verified users is much less than the non-verified users, which reflected the insufficient verified user data and caused a failure of hypothesis test about sentiment scores and stock prices (as we mentioned at the beginning of additional analysis).



**Figure 10. Distribution of Average Sentiment Score by Follower Scales**

This is a graph represents the distribution between number of followers and average of score for Twitter. As the number of followers becomes higher, the score goes lower. In other words, it means the users with large amount of followers tend to be more neutral regarding the stock trend.

***Conclusion***  
 The result of prediction accuracy was 50% before machine learning process. After applying machine learning process, the accuracy increased to 67%, which is a very high accuracy. Therefore we can conclude that Twitter, as a social media, can predict the stock price. To explain this in detail, there are some restrictions about the twitter sample before the method reaches a high accuracy: The proportion of verified user need to be over 6% of the entire Twitter user sample size.

***Limitation***  
 There are a few drawbacks in the project that could be improved in the future study. First, the tickers are abbreviations, which could lead to a misunderstanding and result in inaccuracy. For example, #AAP could stand for American Academy of Pediatrics instead of Advance Auto Parts inc. Second, programs do not have sense of humor. The word count and score may not deliver the true meaning of the Twitter user. Especially nowadays people use sarcastic tones in tweets. The language processing method is also not perfect. By central limit theorem, the score of stocks should follow a normal distribution.   
 Another limitation that raises attention is that the collected tweets include many junk tweets and ads, which are not producing meaningful sentiments scores. For the future study, the result could be improved if more meaningful data is collected.