Using Social Media Posts to Predict Stock Price

DN & Drexel’s Big Data

**Data Science Capstone Project   
Exploratory Data Analytics Report**

Date:

[6/2/2021]

Team Members: Nelson Blickman, Dhruvit Naik

Name: Nelson Blickman

Name: Dhruvit Naik

Name:

Name:

[The purpose of this report is to describe the exploratory data analytics. It includes five major sections:

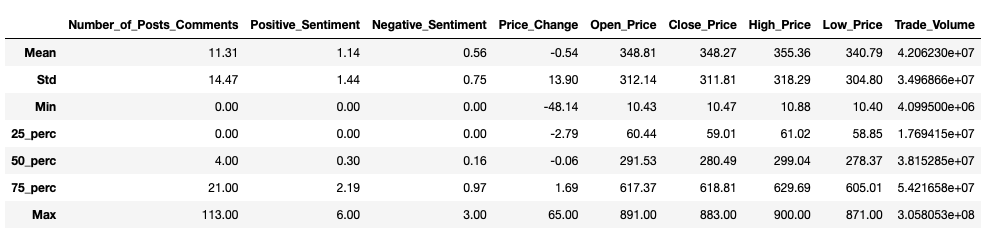
1. Analyzing the basic metrics of variables: data types, size, descriptive statistics
2. Non-graphical and graphical univariate analysis: identifying unique value and counts, histogram, box plots, etc.
3. Missing value analysis and outlier analysis
4. Feature engineering and analysis: correlation analysis, dimensionality reduction, deriving new variables
5. Appendix]

**Analysis the basic metrics of variables**

[In this section, we identify all the variables in the dataset and conduct the basic metrics of the variables. What are the data types (numerical/categorical, discrete or continuous, ordinal or nominal) and size? Provide the descriptive statistics of the variables such as mean, standard deviation, min, max, percentiles, etc.]

The below tables list all of the variables and their characteristics. \*\* Note, this is subject to change a bit based on tweaks that will be made to the Sentiment features before the presentation.





**Non-graphical and graphical univariate analysis**

[In this section, we identify the list and number of unique values for each variables and provide the histogram and box plots to understand the distribution of the data.]

\*\* Please note, the EDA analysis below is subject to change a bit, given our sentiment features will be altered soon.

\*\* Please see Jupyter Notebooks for extensive EDA of anything not included here. EDA Phase I shows a break down of the most common words for each individual company on days the stock went up or down. The compiled words from all company’s is saved as a Json object, and is used for analysis in EDA phase 2 report. Below, I have put the most common words from this compiled Json object.

Up Words Down Words

('cpu', 180) ('even', 247)

('even', 172) ('think', 221)

('see', 168) ('much', 216)

('good', 168) ('new', 215)

('think', 166) ('need', 202)

('much', 166) ('go', 200)

('still', 158) ('could', 198)

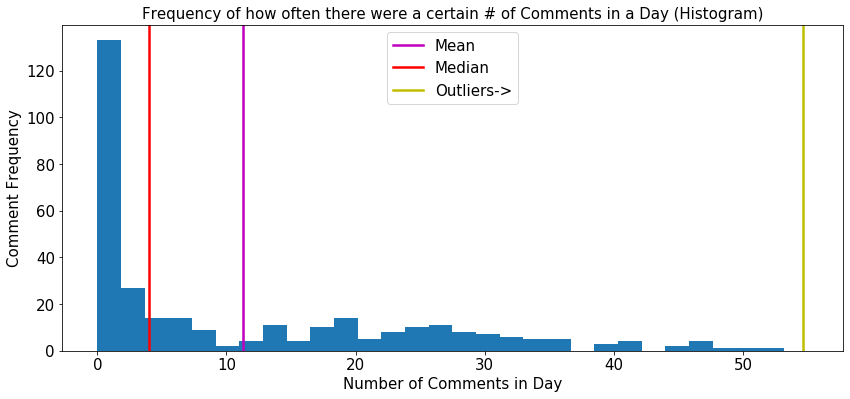
('need', 150) ('buy', 195)

('want', 149) ('good', 194)

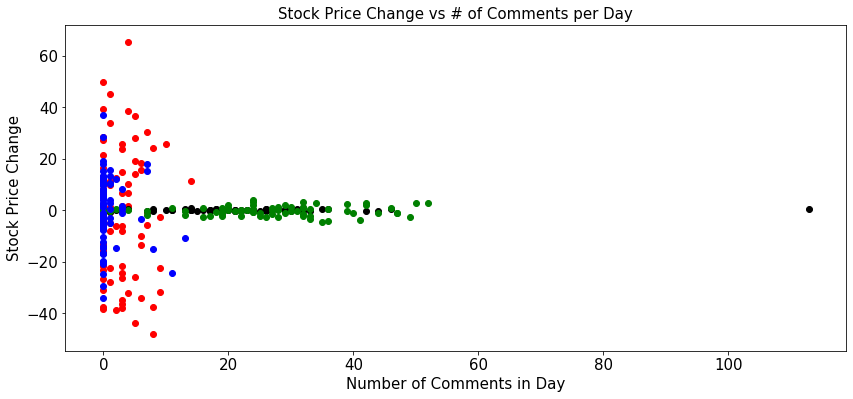
('new', 149) ('cpu', 193)

('use', 146) ('want', 185)

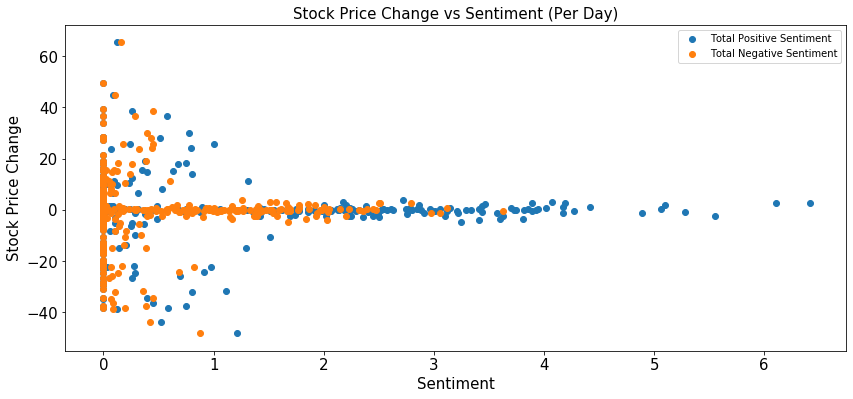
Since each individual record represents the total number of posts/comments on this day, the histogram below was a nice way to understand how often a record had a certain number of posts/comments.



The graph below uses a Scatter Plot to see if there is any correlation between the change in price of a stock on a day, and the number of posts/comments that occurred in a particular day. The four different colors represent the four different companies. It would have been nice to find a clear correlation amongst these variables (stock price vs number of comments), but the distribution on the graph does not show a clear correlation.



Similarly to the graph above, the Scatter Plot below searches for a correlation between movement of the stocks price on a given day, and the sentiment of the posts and comments on the same day. The plot though show’s no clear correlation. Hopefully the complexity of the Machine Learning algorithm that we will implement, which uses much more variables than just these two, will help find correlations in our data.



As mentioned, our compiled word counts which we loaded into EDA phase 2 report were used for other analysis. One of which involved finding the highest word count discrepancies between words used when the stock price go’s up vs. down. For example, the word “even” had the highest discrepancy. This word occurred in days that the stock price went up 75 more times than that word occurred in days that the stock price went down.

('even', 75),

('go', 73),

('could', 69),

('might', 69),

('new', 66),

('since', 58),

('think', 55),

('paint', 55),

('us', 55).

Of course, if the word ‘even’ occurs more often than other words, it would have an unfair advantage in this analysis.Therefore, to improve this analysis, I took the percentage change in the counts below. The word ‘white’ is used in days the stock price went up 87.5% of the time more than on days the stock price went down.

('white', 87.5),

('expedite', 87.5),

('hive', 87.5),

('boy', 86.67),

('batch', 85.71),

('lightning', 85.71),

('beret', 85.71),

('chaos', 84.62),

('bloke', 84.62),

**Missing value analysis and outlier analysis**

[In this section, we identify the missing values and outliers and determine how we handle these values before analysis.]

Our Posts and Comments feature has a few records where there are no posts or comments on a given day, which leads to a couple blank value. However, aside from that, missing values is not an issue in our data. However, we do think we want to get more data. There are no worrisome outliers apparent in the data either.

**Feature engineering and analysis**

[In this section, we identify the variables that are useful for predictive modeling and machine learning through correlation analysis. You may also reduce the dimension or derive new variables so that the predictive modeling can be more efficient and effective.]

As mentioned above, we will tweak the features of our sentiment analysis. However, our variables below seem to all be relevant. We have no plans to remove or add other variables at the moment.



**Appendix**

[Provide the code or pseudo code, and any other information in the appendix here.]

No other applicable code needed at the time. All code will be pushed to Github Repo.

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Analysis the basic metrics of variables** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |
| **2** | **Non-graphical and graphical univariate analysis** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |
| **3** | **Feature engineering and analysis** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |
| **4** | **Appendix** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.