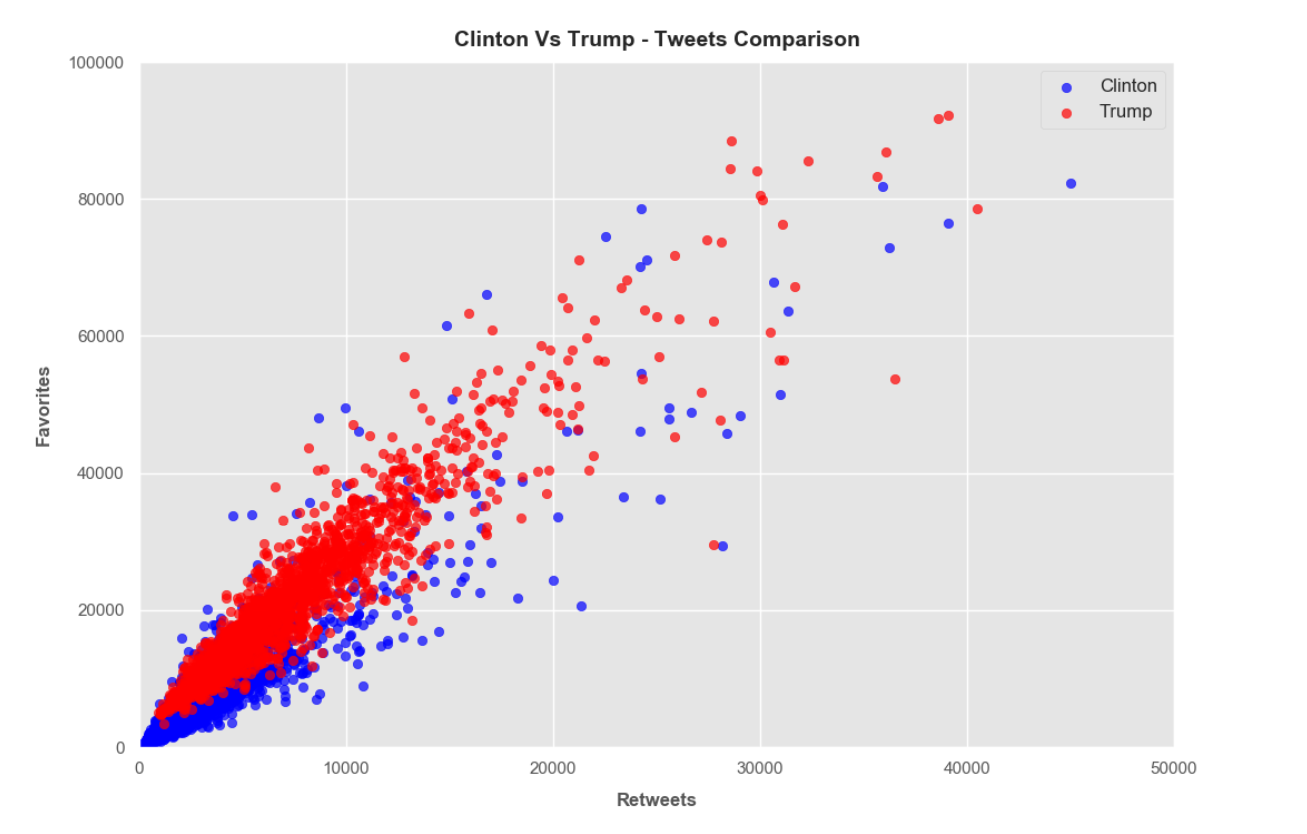
Today, November 8th 2019, exactly three years have passed since the 2016 US presidential election. At this hour back then, the people of the United States exercised their right to elect their president in what turned out to be the most polarizing US election in history. Had I done this twitter posting analysis of Trump and Clinton three years ago before the election, I could have predicted, with significant confidence, that Donald Trump would emerge as the winner.

The figure **17\_ScatterRetweetAndFavorite.png,** shows the comparison of retweets and favorites of the two candidates, as visualized in the scatter chart model.

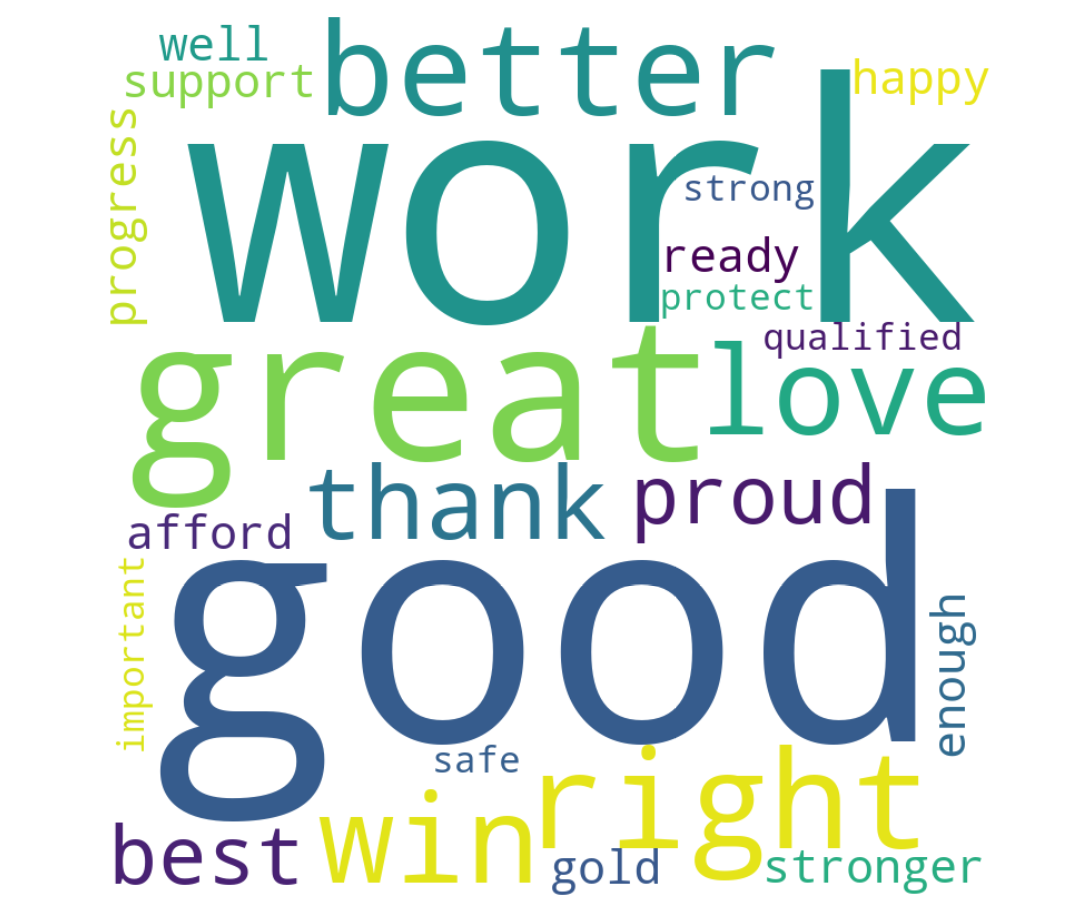
**Fig. 17\_ScatterRetweetAndFavorite.png**

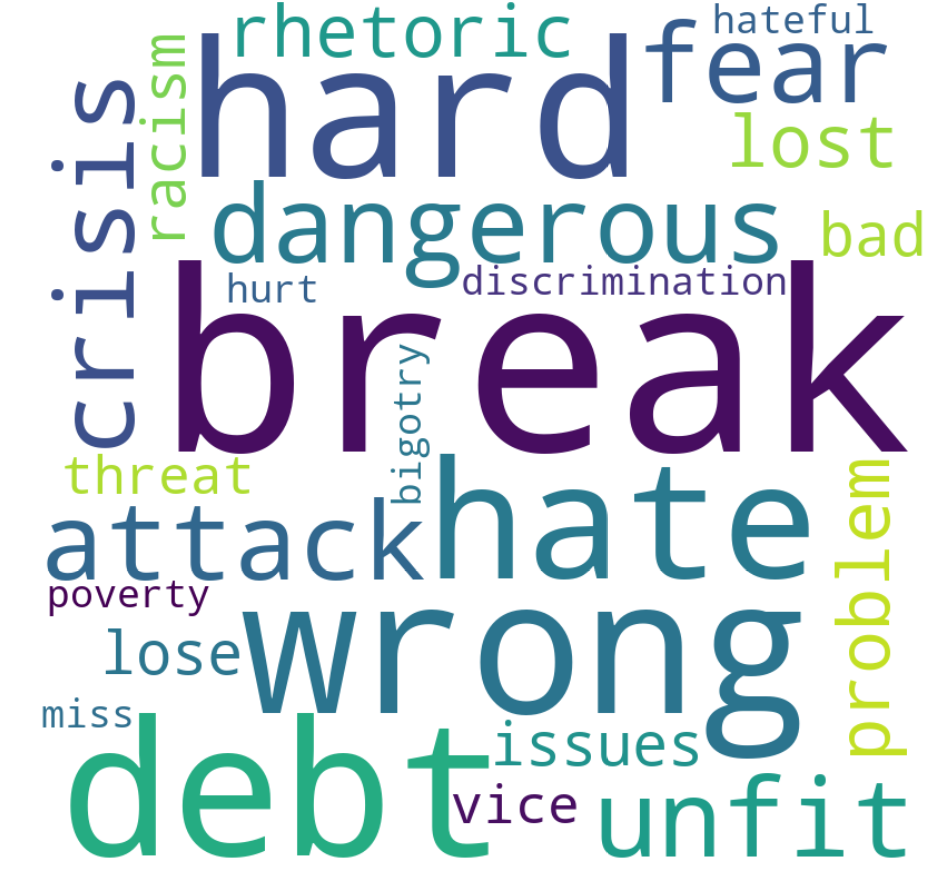


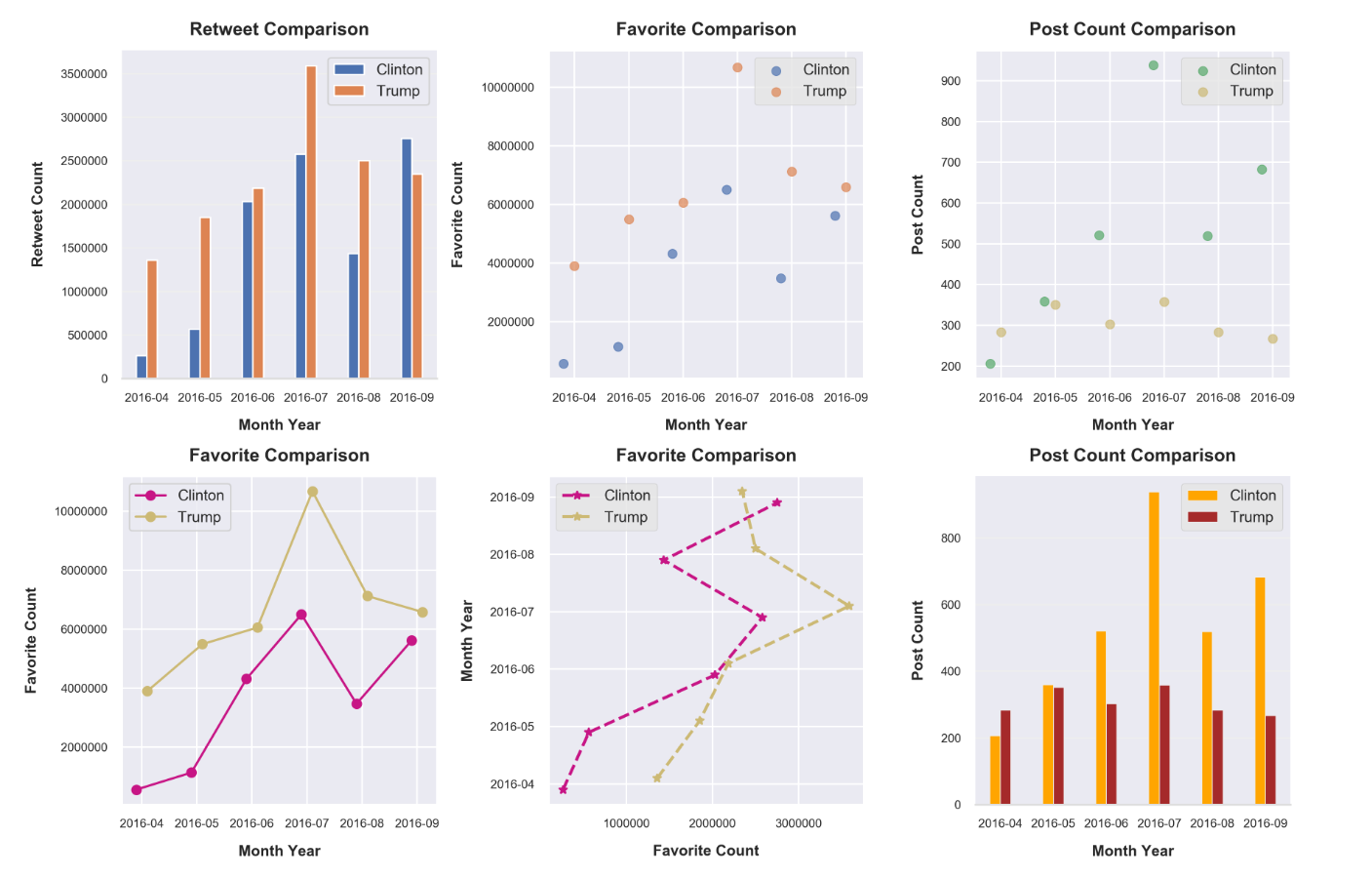
On this chart, one can see the enthusiasm of the Trump supporters in retweeting and liking his posts. Comparing his results to Hillary, the difference can easily be seen, with the scatter chart showing the “big league” discrepancy.

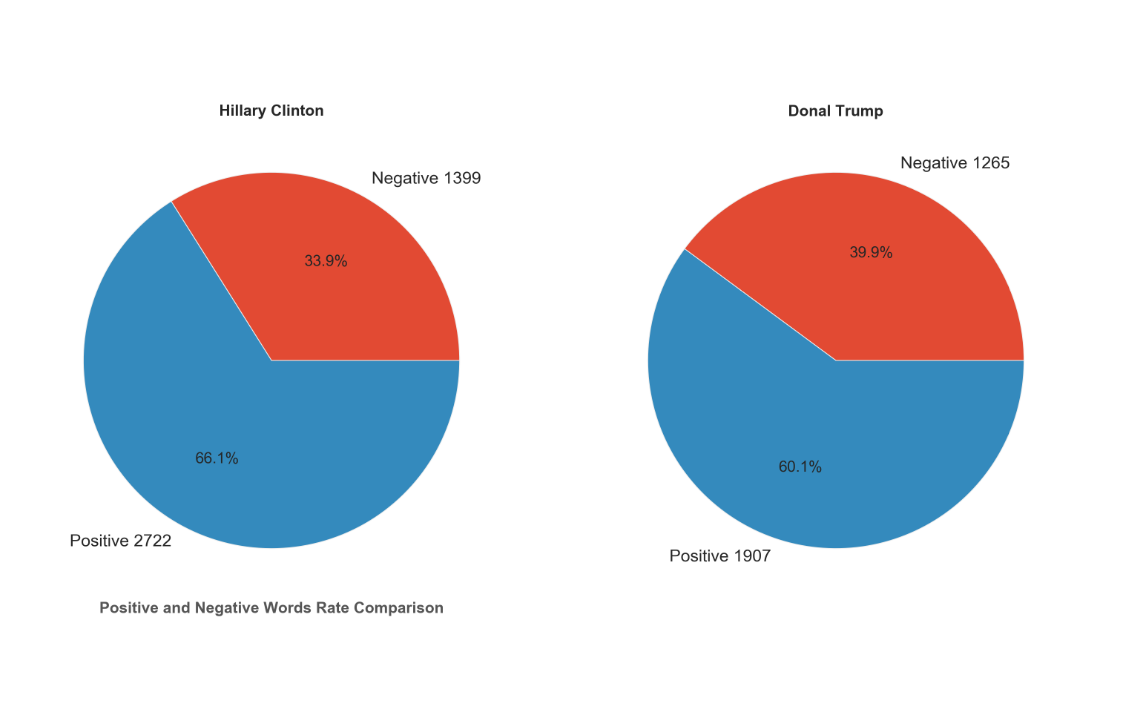
The other charts also demonstrated an easy read of the numbers that indicate that Trump had an underlying advantage. It was also interesting to see the two candidate’s use of certain words, positive or negative, and how they resonated with their respective followers.

**Hillary Clinton Positive Words and Negative Words**

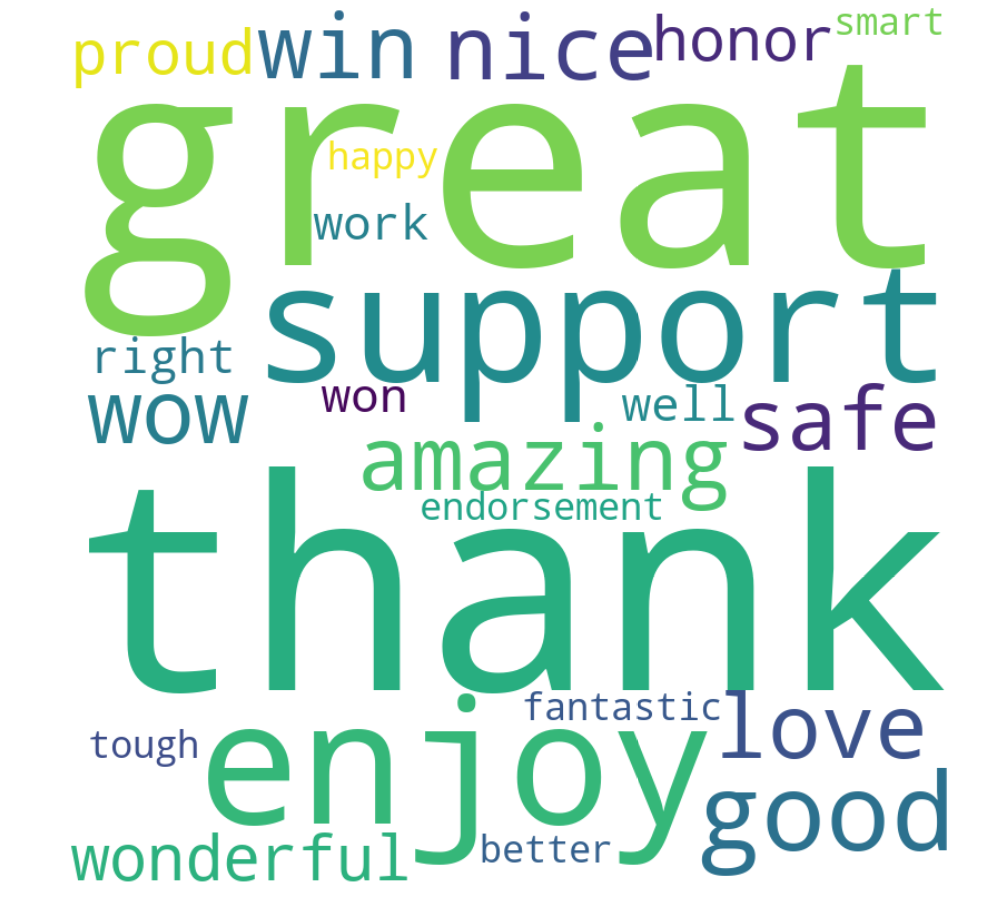




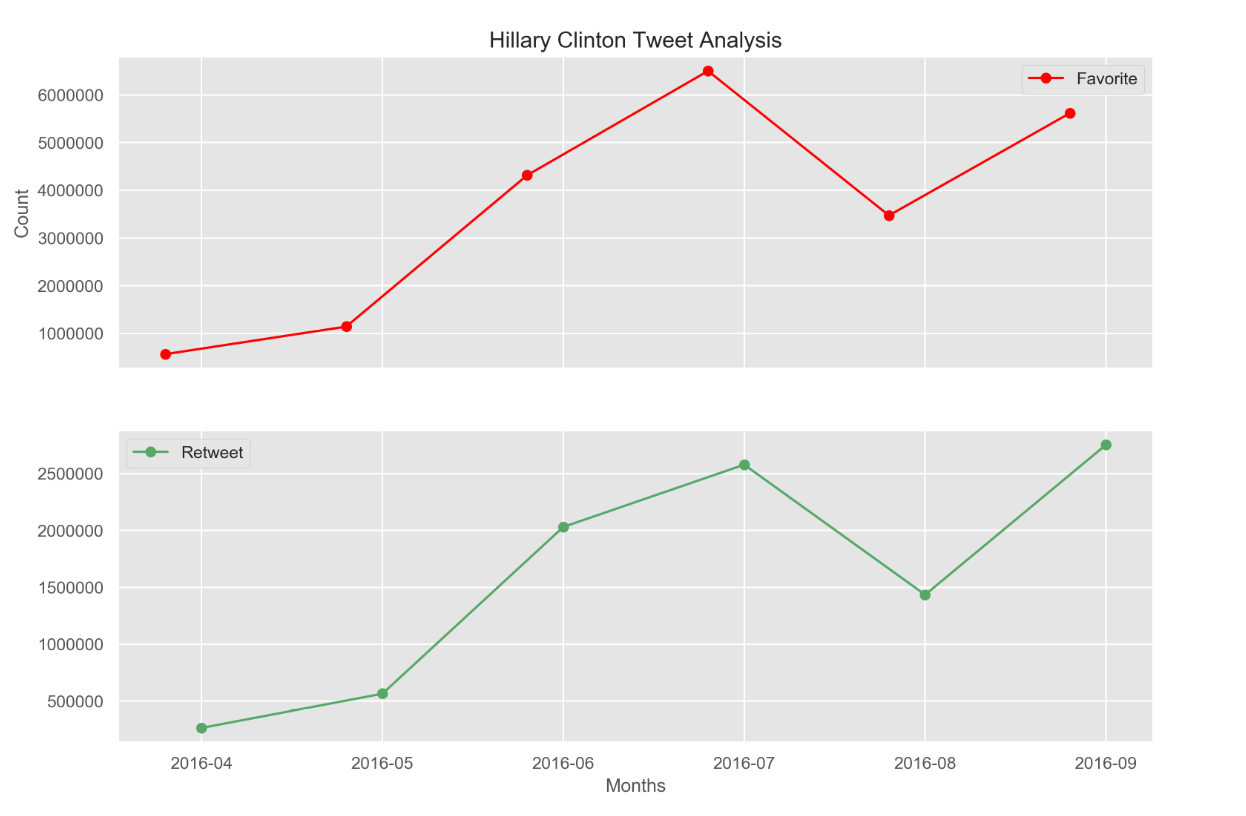


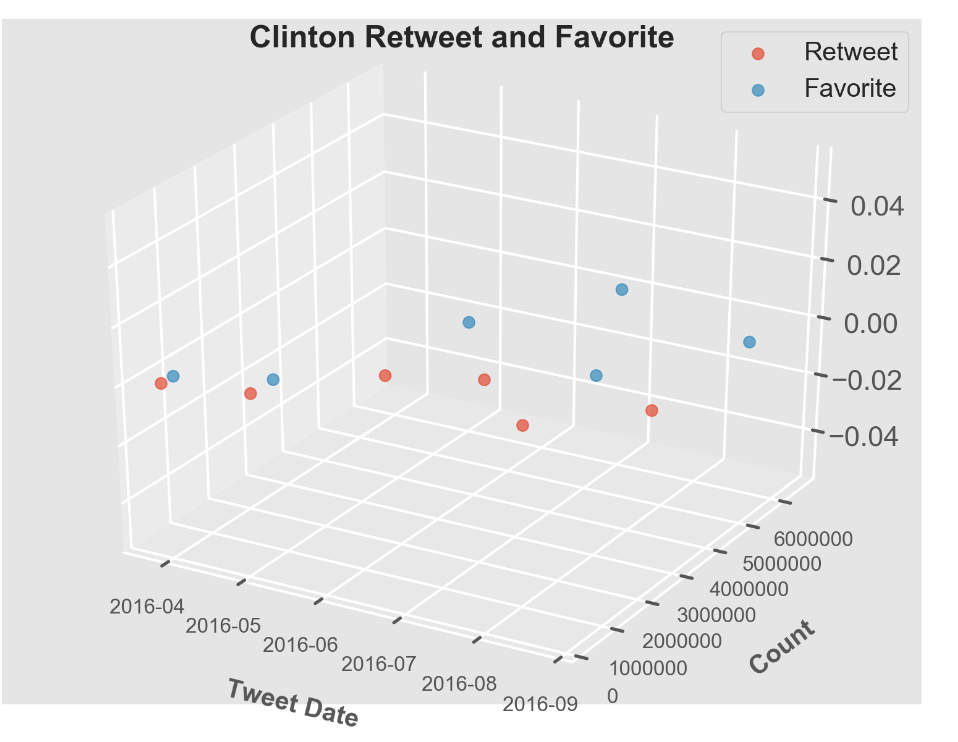


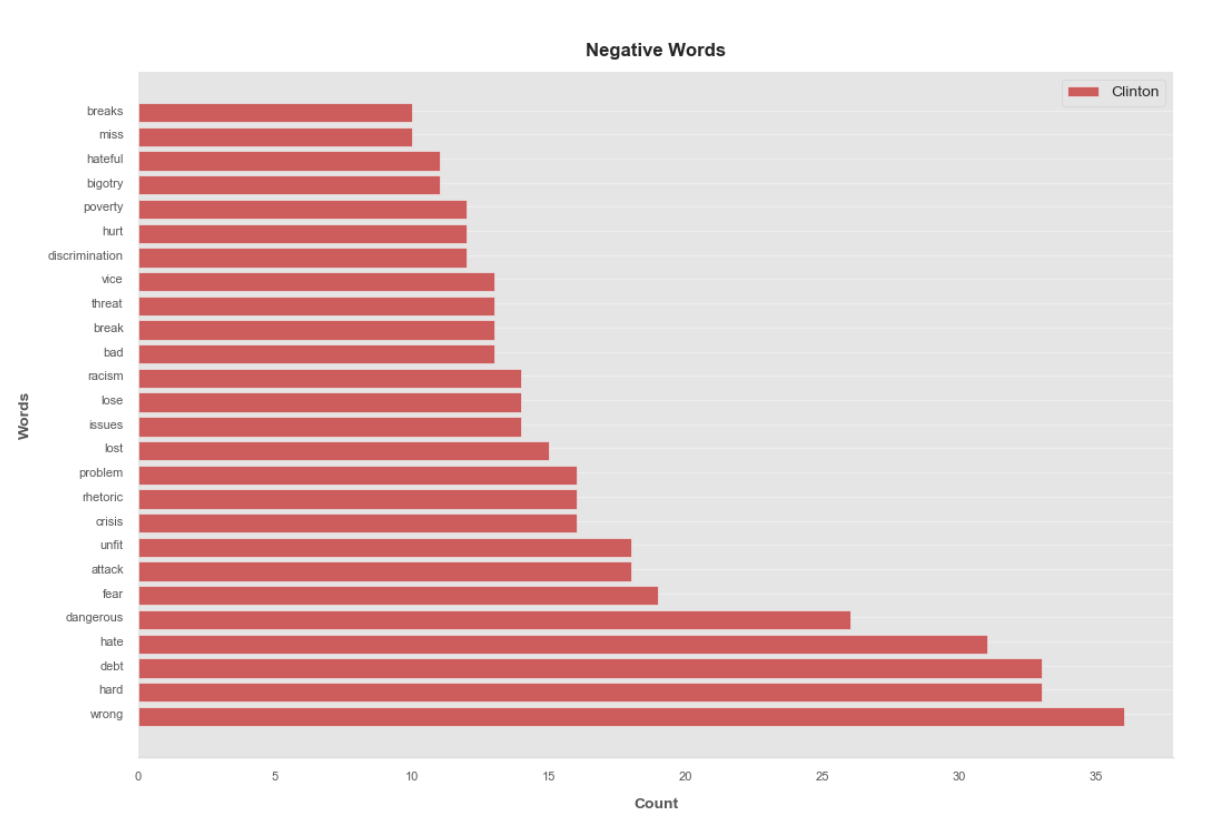
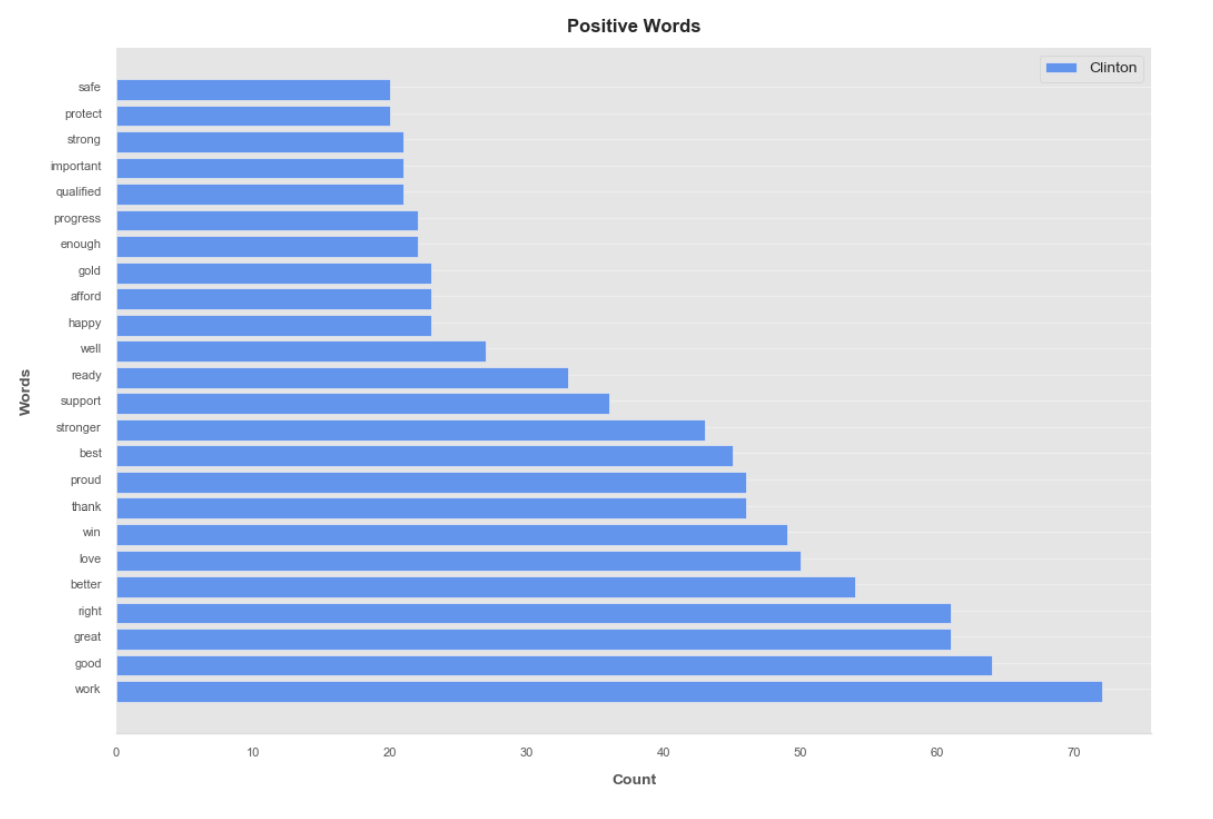
**Donald Trump Positive & Negative Words**

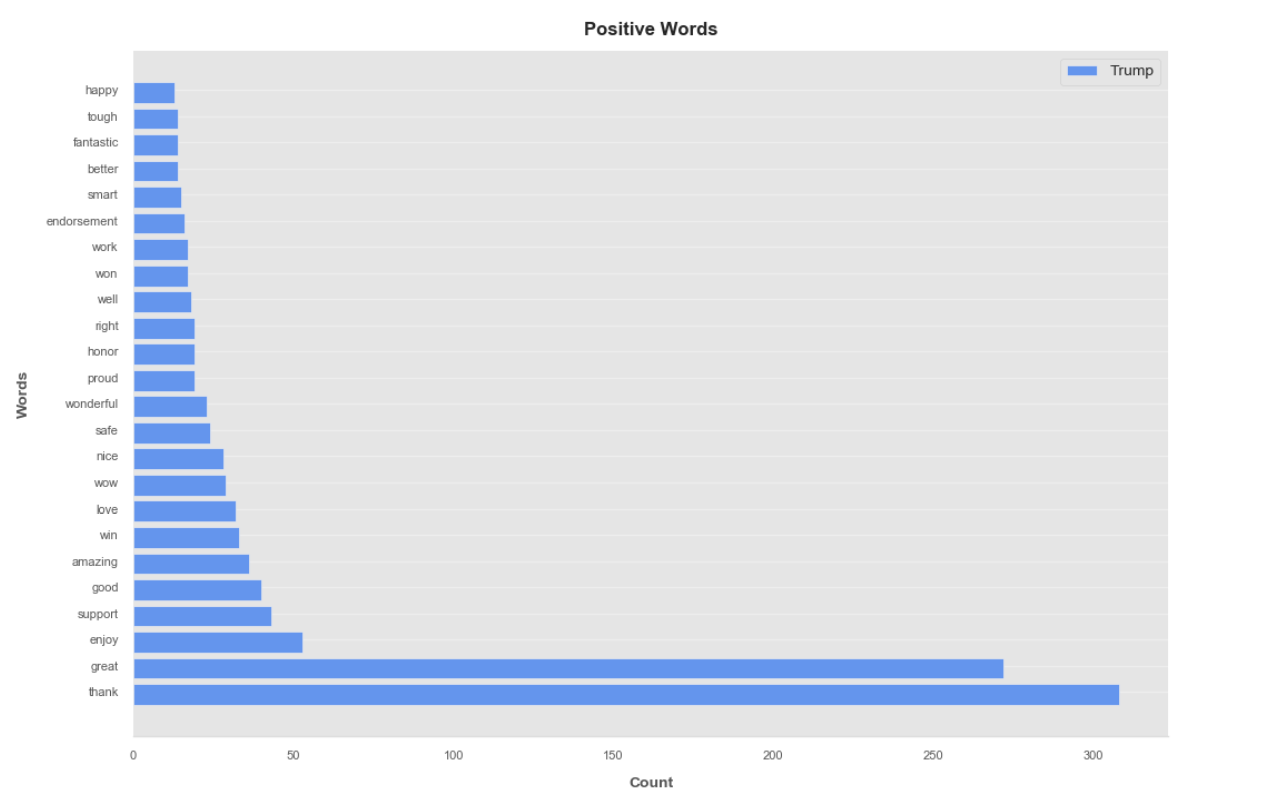


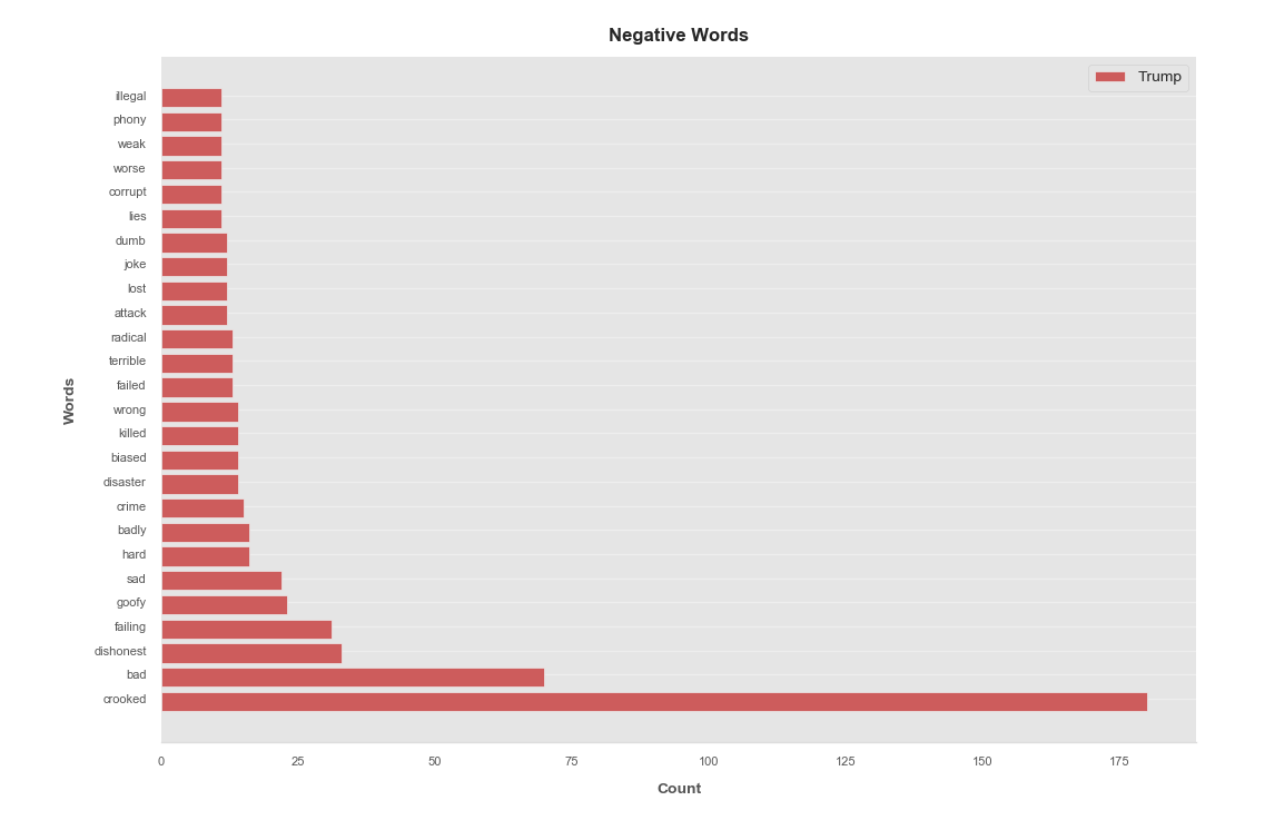












**Clinton Retweet Analysis – Logistic Regression**

**Summary of f1-score:**

|  |  |
| --- | --- |
| **Test Size** | **F1-score** |
| 0.25 | Between 0.66 to 0.69 |
| 0.35 | between 0.67 to 0.71 |
| 0.45 | Between 0.66 to 0.69 |
| I have classified my retweet count as follows:   * Class 1 (1-1000) * Class 2 (1001-2000) * Class 3 (2001-5000) * Class 4 – above 5000   My Features are the following:   * Favorite Count, Hashtag Count, Mention Count, Weekday, Hour, URL Count and Session | |
| **Below is at 0.35 Test Size**  **Accuracy score is 0.69/1**  **Classification Report**  **precision recall f1-score support**  **1 0.71 0.78 0.74 258**  **2 0.64 0.53 0.58 380**  **3 0.67 0.74 0.70 352**  **4 0.81 0.83 0.82 140**  **accuracy 0.69 1130**  **macro avg 0.71 0.72 0.71 1130**  **weighted avg 0.69 0.69 0.69 1130**  **Confusion Matrix**  **[[201 54 3 0]**  **[ 76 203 101 0]**  **[ 5 59 261 27]**  **[ 0 0 24 116]]**  **Overall f1-score**  **0.7130048012347622**0  This is consistently giving me and f1-score between 0.67 to 0.71 | |

**Using KNeighbors**

Using this code below and by changing the n\_neighbors=1 to up t o 5, I noticed that the accuracy value changes:

**Clinton Retweet Analysis – KneighborClassifier**

Performing the KNeighborClassifier using the same dataset, I got the following results:

|  |  |
| --- | --- |
| **Test Size** | **Accuracy** |
| 0.25 | 0.7447335811648079 |
| 0.35 | 0.7053097345132744 |
| 0.45 | 0.7017906336088154 |

Using this code below and by changing the n\_neighbors=1 to up t o 5, I noticed that the accuracy value changes:

clf = neighbors.KNeighborsClassifier(n\_neighbors=1)

clf.fit(X\_train, y\_train)

accuracy = clf.score(X\_test, y\_test)

print(accuracy)

|  |  |
| --- | --- |
| N\_Neighbor = | 0.35 percent test size |
| Without | 0.7061946902654868 |
| 1 | 0.6707964601769911 |
| 2 | 0.6584070796460177 |
| 3 | 0.6973451327433628 |
| 4 | 0.7106194690265487 |
| 5 | 0.7230088495575221 |

The logistic regression is the best fit for my dataset and my objective which is to predict the probability of retweeting a twitter post. With this, I have defined four classifications as the levels of retweet. Then, I have defined the features that I can use as the factors that can predict the outcome.

There are three types of logistic regression namely:

* Binary logistic regression wherein the target variable has only two possible outcomes for example “Pass or Fail”.
* Multinomial logistic regression wherein the target variable has three or more nominal categories for example type of petals.
* Ordinal logistic regression wherein the target variable has three or more ordinal categories for example rating from 1-3.

In my project, I have employed the Ordinal Logistic Regression having 4 classes (1-4) as the retweet number levels.

In the classification report, it shows that my precision is at .7075. It means how often my model prediction is correct. The accuracy is showing as .69 while the F1 is at .71. Based on what I read, the F1 score is better as it considers unbalanced distribution of classes similar to my case.

The confusion matrix is a table that is used to evaluation the performance of the classification model. The diagonal values represents the accurate predictions while the non-diagonals are the incorrect predictions.

**Confusion Matrix**

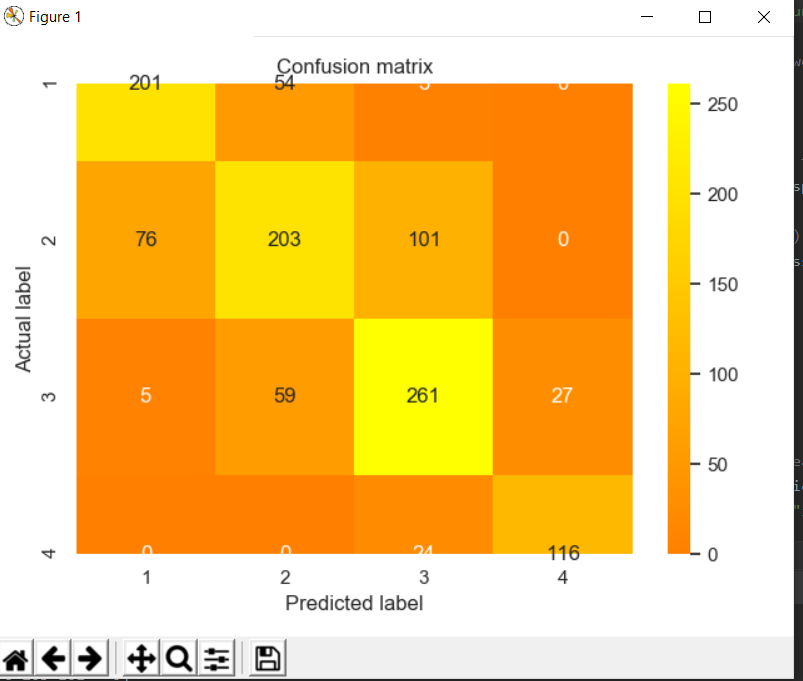
**[[201 54 3 0]**

**[ 76 203 101 0]**

**[ 5 59 261 27]**

**[ 0 0 24 116]]**

Below is the heatmap for my confusion matrix.



I think my model is quite acceptable. Given more time where I can include the “positive and negative words” in the features, I am hoping to get better results.

\*\*\* End of report \*\*\* \*