Project 2B Report

Yutong Han 705025619

June 8, 2018

1 Time Series Plot

The Figure.1 shows the time series plot(by day) of positive and negative sentiment.

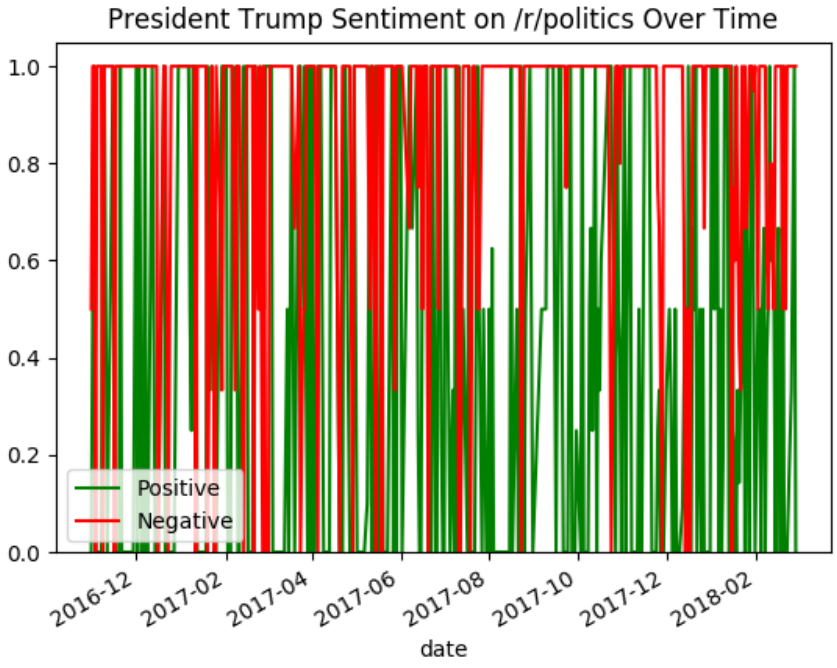

Figure 1: Time Series Plot

2 Positive and Negative Sentiment of US State

The Figure.2 and Figure.3 shows the time series plot(by day) of positive and negative sentiment.

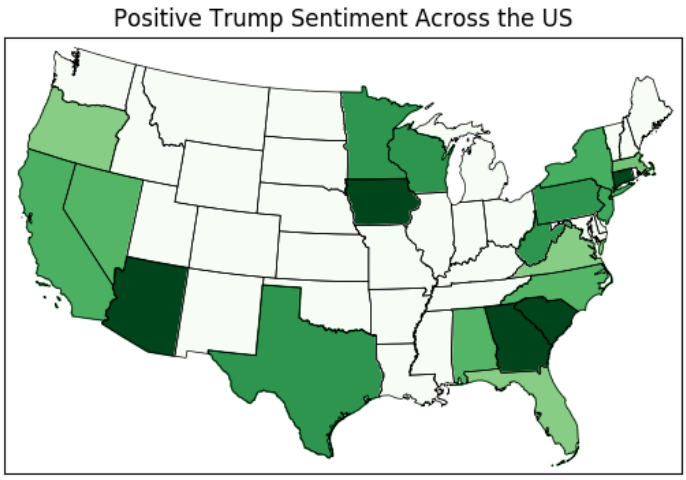

Figure 2: Positive Sentiment over US States

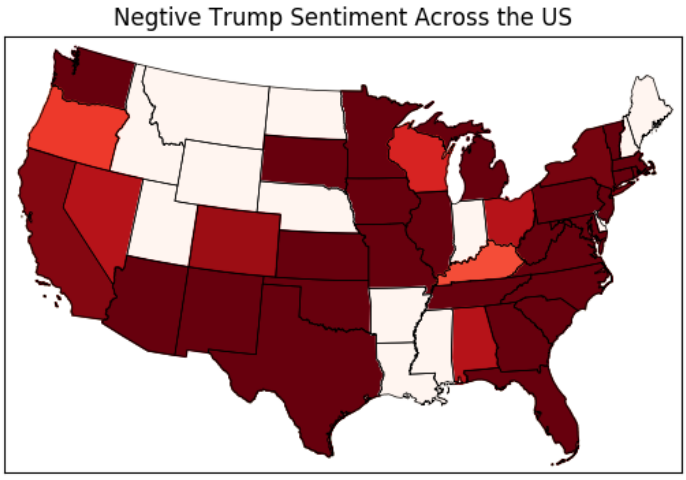

Figure 3: Negative Sentiment over US States

3 Sentiment Difference of US State

The Figure.4 shows the different of positive and negative sentiment across the US States.

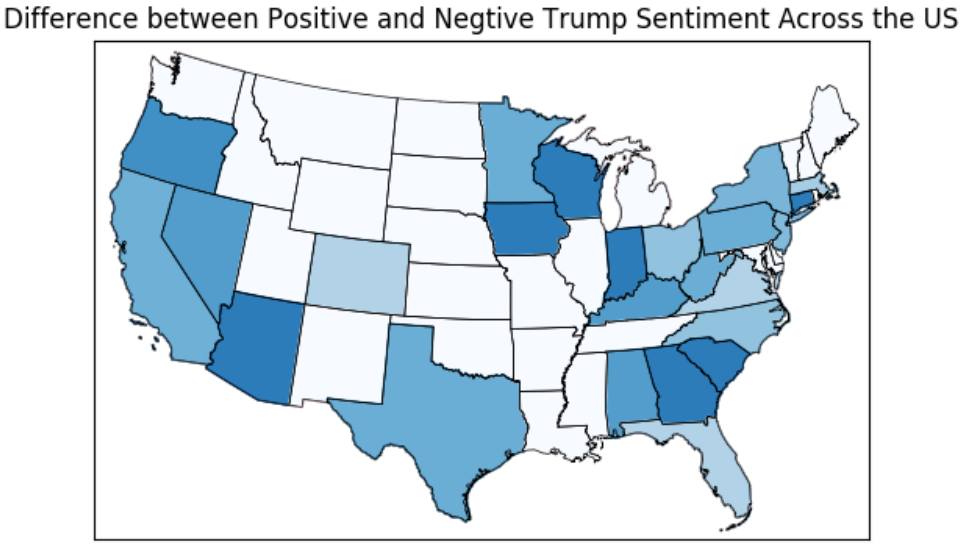

Figure 4: Difference between the positive and negative sentiment across US States.

4 Top 10 positive stories

5 Create TWO scatterplots where the X axis is the submission score,

6 Produce the ROC curves

7 summarizing your finding

8 Question 1

Take a look at *labeled_data.csv* Write the functional dependencies implied by the data.

Answer: The $Input\_id$ is the primary key of the scheme. So the $Input\_id \rightarrow labeldem$, $Input\_id \rightarrow labelgop$, $Input\_id \rightarrow labeldjt$ and their combinations can be implied.

9 Question 2

Take a look at the schema for the comments dataframe. Forget BCNF and 3NF. Does the data frame look normalized? In other words, is the data frame free of redundancies that might affect insert/update integrity? If not, how would we decompose it? Why do you believe the collector of the data stored it in this way?

Answer: It does not look normalized. In the *subreddit_id* and *subreddit* part, it seems that the *subreddit_id* $\rightarrow$ *subreddit*. So the *subreddit* may store repeatedly. So it can be decomposed into two scheme with the *subreddit* be the key of other scheme. Since this database may contains about the posts about the politics, the subreddit is just the */r/politics*

10 Question 3