ISyE 6740 – Summer 2020
Final Report

By Arie Brown (903564701), Gregory Loshkajian (903553788), and Arjun Singh (903073471)
Team #89: Sokol's Disciples

A Journey into Factional Communication: Predicting Party Affiliation from Tweets

Problem Statement

Twitter is one of the world's most popular social media platforms with over 145 million daily active users sending more than 500 million tweets per day. Due to its popularity and concise messaging format, Twitter has also become a hub for politically oriented users to share opinions and reach a wide audience. These politicos come a variety of places, one of the unique features of social media platforms is that television pundits, small time bloggers, and public officials can interact within a truly *public* forum. Many politicians and public officials, most notably President Donald Trump, have recognized this trend, and leveraged Twitter as a new platform to interact with their constituents and potentially expand their base of support.

Because of Twitter's emergence as a platform for political communication, it is possible to view it with the same kind of scrutiny as any other platform. Just as radio and television changed political messaging in the last century, so can politicians utilize Twitter to shape messages, heighten awareness of favored issues, and potentially shift public opinion. With this in mind, our team decided to investigate the usage of language in political tweets by different partisan groups. Specifically, in order to quantify this relationship, we decided to use Twitter data from senators to predict the user's political party.

Using random forest, LASSO, and multinomial Bayes models on a developed feature set based on FiveThirtyEight compiled Twitter data, we were able to achieve an accuracy between ~77% and ~85%. Closer evaluation of the trained models revealed a high emphasis on specific politically charged words and phrases, hashtags, and ideological themes. Some identified themes are very important to Republicans or Democrats specifically. This suggests a political discourse geared towards snappy attacks and easily digestible ideas, as well as a sharp focus on messaging to a base and endorsing fellow partisans, revealing that while political struggle may have adopted Twitter as a new battleground, the platform has influenced the politicians in turn.

Data

The data used throughout this project was obtained from FiveThirtyEight's data repository[1], which was originally compiled with the intent of providing insights on the average twitter "ratio" notable politicians across the political spectrum had. Due to this, the data is not only publicly available, it also has the distinct quality of being well labeled. Each official within

[1] https://github.com/fivethirtyeight/data/tree/master/twitter-ratio

the dataset is marked by party, as well as state. Not only that, the source data also records the number of replies on each tweet, which is incredibly rare to find in public Twitter data, as reply counts are not exposed by Twitter's API. The tweets in this dataset are a collection of 288615 tweets by senators, with history extending from January 2013, to October 2017.

Methodology

Exploration and Feature Creation

Modeling Twitter data presents a set of challenges which are common within the natural language processing literature, for example, how to quantify the presence (of lack thereof) of specific words within unstructured text, what text is worth generating features with, etc. Twitter introduces a suite of other questions, e.g. how should we treat emojis, hashtags, urls, etc? With this in mind, in delivering this project, our team explored the twitter data and identified ways to account for all of these issues, and construct feature suite for a battery of prediction models.

First, to begin preparing the data for the training, we first decided to clean it by removing special characters. Often twitter data contains odd characters introduced by bad encodings. Based on the mistaken characters, it is often possible to recover the true text, which we did. Next, tweets can often contain special twitter specific idioms, e.g. hashtags, urls, and mentions. While hashtags and mentions can be relatively indicative of party, urls are often unhelpful, so we cleaned urls from our corpus. Finally, we dropped common English stop words (e.g. and, the), which are generally not useful or indicative of any particular trait of the text itself.

After our data was cleaned, we aimed to understand activity by user as well as by political party, as this might inform feature construction. We also calculated 2 metrics – positive and negative engagement – in order to better understand the distribution of the tweets. Positive engagement was defined as the ratio of 'favorites' to 'total number of tweets'. Negative engagement was defined as the ratio of 'replies' to 'retweets', under the assumption that a user replies to show displeasure with a tweet, and retweets to signal a endorsement of a tweet. Table 1 shows users with the most tweets and the respective attributes of their tweets while Table 2 consolidates the data by political party.

User	Replies	Retweets	Favorites	Number of Tweets	Positive Engagement	Negative Engagement
SenatorFischer	16,062	23,633	41,794	3,247	12.87	0.68
SenatorTomUdall	27,366	232,065	411,725	3,247	126.80	0.12
Sen_JoeManchin	37,673	48,448	139,521	3,247	42.97	0.78
SenatorEnzi	16,102	24,640	47,828	3,246	14.73	0.65
SenTedCruz	209,229	1,394,307	2,810,555	3,246	865.85	0.15
…						
SenSasse	19,240	57,522	117,488	823	142.76	0.33
SenMurphyOffice	469	1,021	1,349	796	1.69	0.46
SenBillNelson	12,858	25,838	57,565	696	82.71	0.50
SenJohnKennedy	11,532	22,679	83,661	562	148.86	0.51
SenatorStrange	35,647	69,528	274,157	454	603.87	0.51
Total	**12,093,476**	**71,855,961**	**169,255,486**	**288,615**	**586.44**	**0.17**

Table 1: Most and Least Active Senators on Twitter

Political Party	Replies	Retweets	Favorites	Number of Tweets	Positive Engagement	Negative Engagement
D	5,329,723	45,203,541	100,601,212	136,082	739.27	0.12
I	1,376,704	15,183,455	40,006,616	6,255	6395.94	0.09
R	5,387,049	11,468,965	28,647,658	146,278	195.84	0.47
Total	**12,093,476**	**71,855,961**	**169,255,486**	**288,615**	**586.44**	**0.17**

Table 2: Tweets by Political Party

From initial observations, it appeared that the positive and negative ratios were spread out when analyzed on a per user basis. However, when viewed from a political party perspective, we found that Democratic tweets had a much higher (~3.77 times) positive engagement as compared to Republican tweets. The Democratic tweets also had a much lower negative engagement ratio as compared to the Republican tweets. The Independents had a fairly low number of tweets (~2% of entire dataset). This suggests that not only do the Democrats and Republicans have different bases, the Independents look very much like Democrats.

To account for the two distinct distributions we identified in the provided corpus, we mapped both Independents to be Democrats. Since the 2 independents both caucus with the Democrats (e.g. Bernie Sanders) and share traits with the Democratic corpus we feel like this will improve the classification rate of our ultimate model without losing any integrity of the data.

Next, a key feature used in tweets are hashtags. Hashtags often represent themes, campaigns, and ideas, and therefore intuitively seem to be a very effective political tool. To test this theory, we extracted the hashtags in order to identify possible correlations between the hashtags and the political parties. Figure 2 shows the most commonly used hashtags, grouped by party.

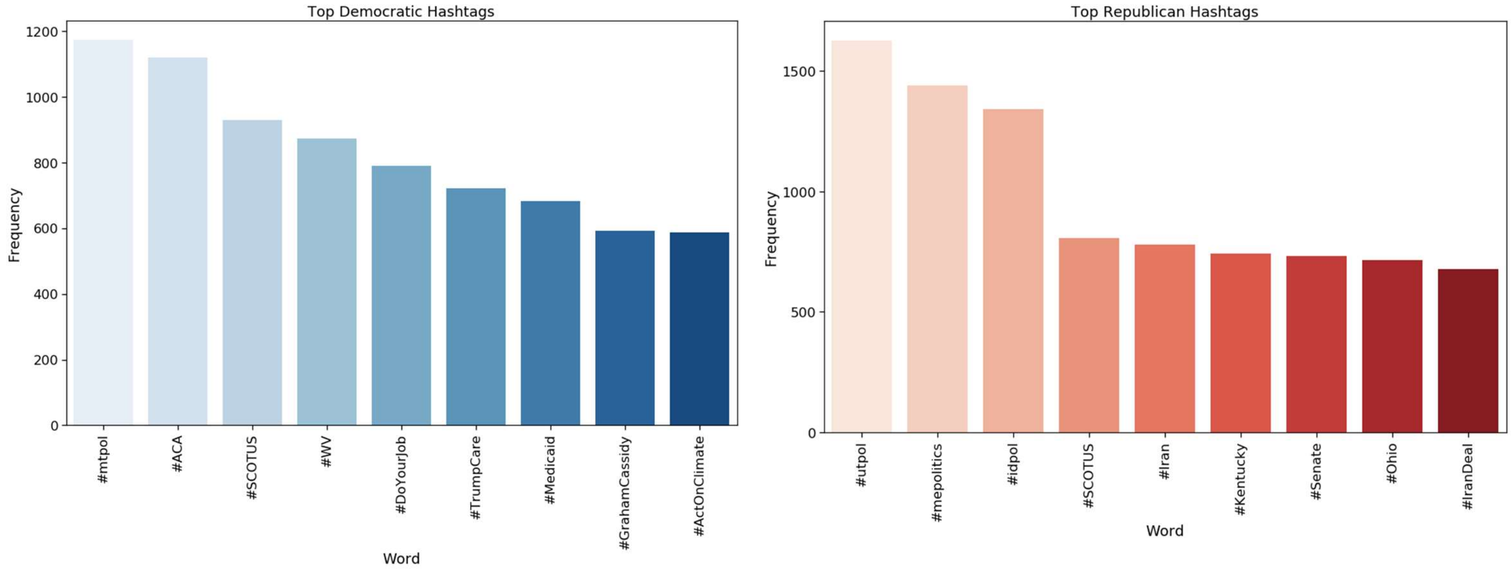

Figure 2: Top Hashtags by Political Party

As we can see above, the Democrats and Republicans use very different hashtags, and therefore, it is reasonably, we maintain hashtags (as well as mentions) as separate terms within our corpus, as these terms will likely have predictive power.

Finally, we transformed our corpus into a matrix representation for training. To do this, we took the top 10000 words from our corpus (which as mentioned previously, had English stop-words removed), and performed tf-idf vectorization on this subset. This approach allowed us to give lower importance to common English words, as well as common political terms (e.g. USCapital), and give a higher weight to words that are novel and therefore may be indicative of specific partisan messaging.

Model Selection

With our feature matrix designed, we ran 3 classification models on our data:

1. Multinomial Bayes
2. Random Forest
3. LASSO Regression, which reached its best performance at $\lambda = 0.25$

In these models, 0 was encoded as Republican, and 1 was encoded as Democrat. Each model was trained using 80% of our dataset, and evaluated on a test set comprised of the remaining 20% of the data. As these models fit a binary classification problem, we chose F1 score as our main evaluation metric. Given that the output we need is a binary flag, the above models were chosen due to their historical strength with classification problems. Alongside this, since this particular problem focuses on a sparse dataset with many features, we selected models which could potentially perform well within a large feature space.

Final Results

Upon evaluation, the three training models listed above had the following prediction F1 scores, 83.9% for Bayes, 76.7% for Random Forest, and 85.3% for Lasso. Figure 3 below shows confusion matrices for each of these models. Observations on these results follow:

1. Random Forest performed the worst. A future exercise might involve re-running and forcing the algorithm to include the favorite and ratio features. As is, picking a random set of words without the guarantee of those normalizing features led to poor prediction accuracy. (Republicans and Democrats largely talk about similar topics and tweets from there can be differentiated by how they are perceived.) Interestingly, the confusion matrix for Random Forest shows that the model has a large number of false negatives, which points to Republican tweets being confused for Democratic tweets.
2. Multinomial Bayes did relatively well on this data, which is reasonable, as this model is well specified to handle sparser data matrices like bags of words.
3. LASSO performed the best which is likely due to its optimization procedure favoring sparse matrices. Given the large number of features, including all of them would likely lead to overfitting which tends to decrease prediction accuracy on outside data.

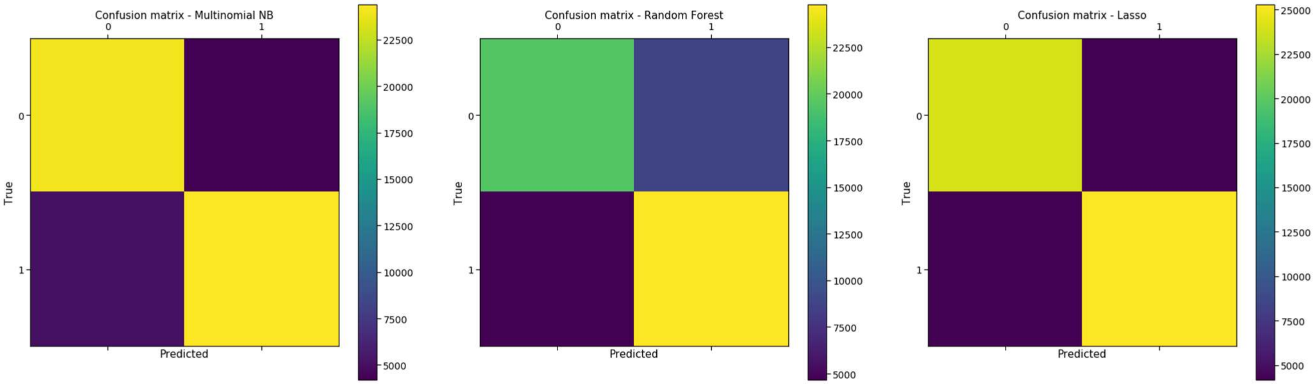

Figure 3: Classifier Confusion Matrices

What can these results tell us? In particular, are our models identifying real distinctions in political communication? To explain this, we can pull the coefficients from LASSO, our best performing model. By checking the coefficients LASSO prioritized, we can identify which words have strong enough prevalence to avoid being shrunk to 0.

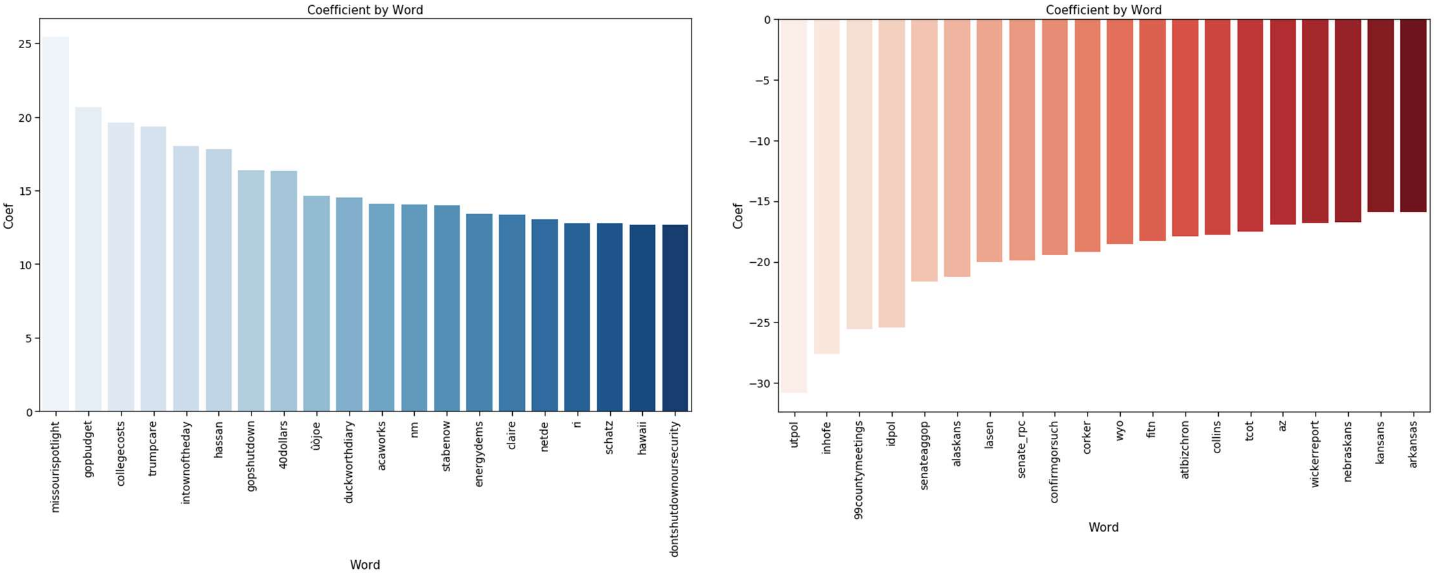

Figure 4: Top Words by Coefficient

As we can see in Figure 4, the LASSO identified influential words are rather clear, and often describe specific political events. For example, "acaworks", "trumpcare", "gopshutdown" all have positive coefficients, which mean the presence of these words in a tweet suggest the tweet is likely to be Democratic. As it turns out, each succinctly displays a liberal objection to some Republican policy. Politician names are also relatively common, suggesting that politicians often refer to politicians within their own party, usually to signal support. With this in mind, it appears that while politicians have brough some habits, e.g. the storied traditions of mudslinging and endorsements, Twitter has shaped these maneuvers to be concise and easily digestible, in other words, very *tweetable*.

Future Ideas/Further exploration

In the future, this analysis could most be improved by doing more data cleansing especially related to the tweet data. For instance, we may have achieved even stronger results by marking some hashtags as generally conservative or liberal, or by using the presence of particular named entities within a tweet (as opposed to simply twitter mentions). Additionally, there are a few ways in which this analysis could be extended.

1. Apply this analysis to non-senator twitter users, such as cabinet members, congresspeople, and presidents. As the model does not currently consider the levels of engagement, it might also be possible to have this model predict party affiliation of other groups of twitter users as well. Given the impending 2020 elections, it might be valuable to use such a model to see where the general Twitter audience seems to be leaning politically, as well.
2. In addition to the analysis presented in this report, one can also extend to identifying the political leaning of a person’s twitter network based on their tweets and the reaction to those tweets.

Collaboration

In the course of taking on this project, each member of the team focused on different aspects of the modeling process. This separation of duties is outlined below, for the reader’s awareness.

Task	Members involved
Project Proposal	Gregory
Initial preprocessing, EDA, multinomial Bayes evaluation	Arjun
Feature extraction/Random forest/Lasso evaluation	Gregory, Arie
Analyzing results	All
Report	All