PREDICTING POLITICAL PARTIES THROUGH TWITTER AND MACHINE LEARNING

Matthew Antle, Chase Dooley, Julia Lanzel, and Timothy Urness
Department of Mathematics and Computer Science
Drake University
Des Moines, IA

ABSTRACT

With the rise in the use of social media outlets by many politicians throughout the United States, many researchers and pollsters are turning to Twitter to track trends among both politicians and US citizens to predict the future of politics in this country. When looking at national politicians of both the House of Representatives and the Senate, their tweets contain both personal remarks, as well as opinions about very heated topics in today's society. This research involves looking at Congress men and women's filtered tweets that took place at any point within the last year and determining, through machine learning and use of neural networks, if their political party could be predicted on that basis only. By running both training and testing data through the machine learning algorithms, it was concluded that this method was fairly accurate, and general trends persisted, even as more epochs were added. This work demonstrates the future of machine learning in the political polling process of any politician, as well as the use of machine learning to analyze data from Twitter in general.

INTRODUCTION

Two years ago, Donald Trump took the world by storm with his presidential campaign. Media outlets were constantly informing the public of what Trump had said that day, or rather, what he had tweeted that day. Trump used Twitter to share his platform and political agenda with the rest of the world, and while some may not have agreed with this form of communication for serious business, experts agree that Trump helped modernize the forms of communication used, just like previous presidents have done [4]. Politicians across the board keep up with Trump's use of Twitter, and in the past years, both Democratic and Republican members of the House and the Senate tweet regularly about their own platforms and political agendas.

Background

Congress has come a long way in terms of how it communicates with the general public, all starting back in 2006, when Twitter made an entrance into the culture of social media [5]. Ever since 2006, Twitter has been gaining international popularity, and its official entrance into the world of political communications took place during the 2008 election through Barack Obama's campaign [7]. While Obama began his presidency, Republicans began to question what went wrong and how they should move forward. In 2009, John Boehner, the Minority Leader of the House of Representatives at the time, pushed Republicans to join Twitter to surprise and surpass Democrats use of the communication platform [5]. When Boehner made this announcement during February 2009, only 69 Congress men and women were presently on Twitter, but by September 2009, 159 Congress men and women held accounts on this social media platform [3]. Twitter allowed Congress men and women to communicate easier with younger generations, but it also appealed to them since it allowed them to run cheaper campaigns. According to Lassen and Brown in their 2010 article on the current political communications, Twitter allowed Congresspeople to make short, descriptive comments about their current platforms without having to run television commercials. Even though the tweets might not have received the same viewer amounts, the press picked up and reported back on the changes in their platforms, providing more coverage than the tweets itself. By 2010, 10 out of 12 Republicans in the house were on Twitter, compared to only 1 out of 17 Democrats in the House [5]. However, a 2012 study done by political scientists at the University of Southern California showed that by the end of 2011, over two-thirds of Congress men and women were now on Twitter, a drastic change since early 2009. By the time the 113th Congress entered the Capital Building on January 3rd, 2013, all members of Congress held at least one Twitter account, if not separate accounts just for their campaigns [7]. When Donald Trump began his campaign in 2015, Twitter had made itself a thriving form of communication on the political scene.

Twitter use remains popular among political icons, but their Twitter usage strays away from how most users use Twitter. In a 2009 study done by the University of Maryland College of Information Studies, the main uses on a Congress member's Twitter account included blog posts, web links, or information on upcoming events, whether they are political or personal events. Another 2009 study reported by the Congressional Research Service, researchers kept track of six congressional activities on Twitter: position taking, press or web links, district or state activities, official congressional action, personal posts, and replies to other tweets. This study found that the most common use of Twitter by Congress men and women included press or web links, which made up 43% for in session and 46% during recess periods. Many of these press or web links referenced political articles or their own blogs, giving them their own press coverage to get the general public involved in their own platforms. This approach makes sense, considering by this point, nearly half of Americans between the ages of 34 and 55 were using Twitter and other social media as their main source for national news [6]. In terms of content, by 2013, Democratic Congress men and women posted mainly on marginalized groups, whereas Republican Congress members were more likely to post about health care [8]. The language itself is also heavily scrutinized, and in 2015, Republicans were 4% more likely to use negative rhetoric than their Democratic counterparts [7]. In 2016, politicians across the board used words like "Tax" and "Black" the most, and as the year went on, the words "Email" and "ISIS" began very popular as well [2]. In the last ten years, gigantic amounts of data have been collected on how Congress people are using Twitter, but recent studies have been reaching out to do more work on the actual sentiment behind the tweets themselves. With this in mind, the aim of our study is to use machine learning and algorithms to sort tweets as Democratic or Republican based on the sentiment of the language being used.

Related Work

Recent studies have focused on the content of the tweets that members of Congress have made, including the sentiment behind the rhetoric. A study down by Annelise Russell out of University of Texas, Austin focused on categorizing the tweets during the first six months of the 113th and 114th Congress, and she collected 33,830 tweets from January 3rd, 2013 through June 30, 2013 as well as 55,235 tweets from January 3rd, 2015 through June 30, 2015. She analyzed whether tweets contained positive, negative or neutral rhetoric, as well as partisan versus nonpartisan rhetoric through manually coding tweets before running them through a machine learning system. Russell eventually discovered that 17% of all Republicans tweeted partisan tweets during the 2013 session, with over two-thirds of these tweets being negative, direct attacks of the democratic party. On the other side, only 5% of Democratic tweets were discovered to be partisan during the 2013 session, with 50% being direct attacks on the Republican party [7]. Overall, Russell used coding and machine learning to weight the level of partisanship and positivity among the Congress member's tweets.

Another study done by Delenn Chin, Anna Zappone, and Jessica Zhao out of Stanford University tested three different algorithms to figure out the emotions behind the emojis of tweets: Support Vector Machine (SVM), Naïve Bayes, and Nearest Neighbors. This Stanford Research Group worked with 2,000 tweets without emojis to train the data for each of the three algorithms before entering 3,000 of the emoji-filled tweets through the machine learning system. Before entering the tweets into the system, the researchers stripped the tweets of any punctuation, URLs, and common words that are unrelated to the sentiment, and later compared them to the same tweets with all of those included to conclude that by taking out the punctuation and common words, classification was made easier, whereas by keeping the URLs, classification was made much more clearer. When using the three algorithms, SVM was proven to be the most accurate, classifying 49.22% of tweets correctly, with the Naïve Bayes following close behind by classifying 49.02% of the tweets correctly [1]. While this study helped with understanding the need to manually enter tweets and which algorithms were most effectives, the results among the emojis seemed biased considering most tweets were based in California, Florida, Texas, and New York.

The final study we looked focused on positive to negative sentiment of tweets during the Presidential debates, and this study came from the University of Pittsburgh with researchers Tianyu Ding, Junyi Deng, Jingting Li, and Yu-Ru Lin. Researchers in this study worked with the Naïve Bayes, classifying tweets based on a zero to one scale, with 0 = Positive Sentiment and 1 = Negative Sentiment. Through this study, researchers used 12,168 tweets from debate one, 11,204 tweets from debate two, and 5,324 tweets from debate three [2]. To fully analyze the sentiment of the tweets, researchers started by calculating the sentiment score from a lexicon designed by researcher Hu and Liu in their own research study before doing and Gaussian and Bernoulli Naïve Bayes classified. Before the tweets were run through the classifier, all hashtags, "@" symbols and anything following them, retweets, and the names of the candidates were removed, and 187 of the tweets were manually classified (70 positive and 117 negative [2]. This University of Pittsburgh study concluded that the Naïve Bayes method had a 75% accuracy rate with the conditions were applied. Overall, these studies helped show which algorithms were more effective for our machine learning system, why some tweets need to be manually entered, and the importance of keeping certain information from the tweets and removing other information.

METHODOLOGY

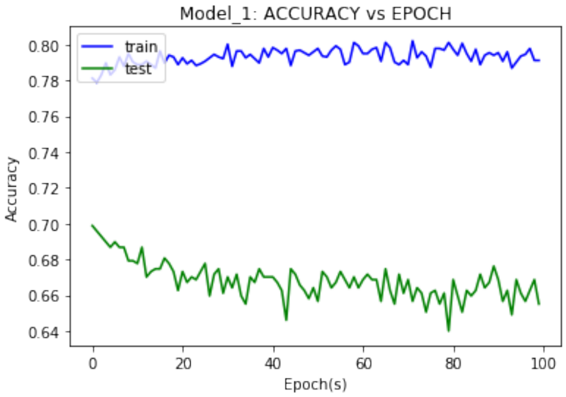
In order to get started collecting tweets, we separated the tweets into two groups: one for training data and one for testing data. The training data is the data we will use to eventually train the machine learning algorithm we created, and the testing data will eventually go through the algorithm itself. The tweets from the dataset were manually collected from recent twitter feeds from several Democratic and Republican Congressmen, as well as from VoteSmart's API database [10]. Because the tweets are from National political figures, they are confirmed to be of their respective political bias. Tweets were separated into two text files for both the training and testing datasets, one for Democratic tweets and one for Republican tweets. A corresponding text file was then created for each file to label the tweets. For this analysis, tweets recognized as Democratic were scored as 0 and tweets recognized as Republican were scored as 1, showing that they are on opposite ends of the sentiment analysis spectrum. Using the BeautifulSoup library in Python, all tweets were scraped of their HTML markup. To further the simplification of analyzing the tweets, all hashtags, symbols, URLs, and any remaining non-letters were removed due to their complex nature. For example, to a human, the hashtag '#NoMoreHomework' could easily be interpreted as the phrase 'no more homework.' However, because it would require a more sophisticated algorithm to understand this, all hashtags, as well as other non-words, have been removed.

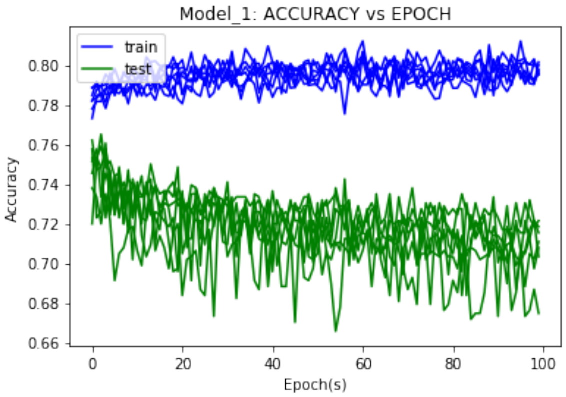
After all this data was filtered, democratic tweets totaled **(DEM TRAINING)** in the training dataset and **(DEM TESTING)** in the testing dataset. For republicans, we trained **(REP TRAINING)** tweets and tested **(REP TESTING)** tweets. Finally, the cleaned tweets were split into lists, with each element in the list representing a separate year.

For the next part of the analysis, Google's Word2Vec, a model that converts normal text into numerical vectors, was utilized. Word2Vec encompasses an expansive list of words and phrases with scores for each [9]. Rather than analyzing the individual vectors of each word, the average vector score was calculated for each tweet by summing the vector score of each element in each tweet and dividing by the total amount of words in the tweet. The output for each tweet was a 1x300 list of vector scores. Next, the data was separated into testing and training sets. Roughly 20% of the raw data was placed into the testing set. The lists were then converted into numpy arrays using the NumPy Python library.

Once the raw data was converted into numeric form, a neural network was trained. Utilizing the Keras API in Python, a sequential neural network with two dense layers was created. A neural network is a machine learning algorithm designed to mimic the way a human thinks. A dense layer is a layer that has all of its neurons connected to all neurons in the previous layer. The first layer contained 300 nodes and each node received each of the 300 vector scores as input. The second layer contained 1 node and utilized the Sigmoid Activation function in order to keep values between 0 and 1. After compiling the model, it was trained on the training tweet vectors scores and their respective political score (either 0 or 1). After the training phase, the testing data was used to evaluate the accuracy of the model. The number of epochs, or times the model went through the training data, was set to 100. To analyze how consistent the model is, the process of randomizing the data was repeated 7 times, and the model was trained on each of the different data sets. To visualize the accuracy of the model on varying data sets, graphs were constructed using the MatPlotLib library.

RESULTS





The left graph visualizes the accuracy of the training and testing data from the model. The minimum accuracy obtained by the training data was 77.8%, while the maximum accuracy was 80.2%. Likewise, the minimum accuracy obtained by the testing data set was 64%, while the maximum accuracy was 69.9%. In general, the model obtained a testing data accuracy in the mid-to-upper 60th percentile.

The right graph visualizes the accuracy of the same model trained on 7 different data sets. According to the graph, the average minimum accuracy of the 7 training data sets is 78%, and the average maximum accuracy is 81%. For the testing data, the average minimum accuracy was 69% while the average maximum accuracy was 76%.

The general trends appear to be consistent across different data sets. However, the range of possible accuracies is greater among the testing data sets compared to the training sets. This may be due to multiple reasons. For one, the tweets used in the training set may vary slightly from those used in the testing data set, thus inhibiting the models ability to predict accurately. This may be caused by the wide spectrum of topics covered in the individual tweets.

Another observation from the visuals is that after the first 2 or 3 epochs, both the training and testing data set accuracies appear to flatten out. Furthermore, during the first 2 or 3 epochs, the training data set appears to slightly increase in accuracy while the testing data set appears to decrease in accuracy. This may be a result of the data being tailored more towards the training set rather than the testing set, which again may be caused by the many topics discussed in the tweets.

CONCLUSIONS

According to our results, our models were able to moderately determine a tweet's political leaning between Democratic and Republican. The results show that while the accuracies of our model's prediction began high, peaking at 76% for the testing data of the repeated model, there is a negative trend as more epochs are used. Furthermore, between the single model and repeated model, the later produces better predictions, which suggests that having more datasets go through the model may produce a better, more accurate result compared to one large set. The difference between the min and max accuracies for our repeated model was 7%, compared to 5.9% in our single model. While the gap grew, the repeated model was more accurate.

Overall, as the number of epochs increased, the testing accuracies decreased while the training accuracies increased. This could be explained by our training model having more data, and so our model may have developed better confidence in its predictions; however, more investigation into this is required for a definite answer. This phenomenon could be caused by a multitude of factors, including overfitting, tailoring, numbers of layers, etc.

Future Work

We wish to emphasize that these results suggest that a correct mixture of epochs, layers, model types, and other factors may exist that will produce more accurate predictions. Investigating the relationship between these factors, and among others, and their effects on the accuracies is valuable. Furthermore, testing these prediction models with other datasets may produce interesting results as well. Future work may want to analyze the predictions themselves, and implications and applications of such results and models. Lastly, as more trials are done with these results, the more accurate the algorithms themselves will become.

REFERENCES

[1] Chin, D., Zappone, A., Zhao, J., Analyzing twitter sentiment of the 2016 presidential candidates, https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf, published January 12, 2018.

[2] Ding, T., Deng, J., Li, J., Lin, Y., Sentiment analysis and political party classification in 2016 U.S. President debates in Twitter, sbp-brims.org/2017/proceedings/papers/ShortPapers/SentimentAnalysis.pdf, accessed January 20, 2018.

[3] Golbeck, J., Grimes, J., Rogers, A., Twitter use by the U.S. congress, Journal of the Association for Information Science and Technology, (61), 1612-1621, 2009

[4] Keith, T., Commander-In-Tweet: Trump's social media use and presidential media avoidance, https://www.npr.org/2016/11/18/502306687/commander-in-tweet-trumps-social-media-use-and-presidential-media-avoidance, published November 18, 2016.
[5] Lassen, D. S., Brown, A. R., Twitter: the electoral connection?, Social Science Computer Review, 29, (4), 2010.
[6] McGee, Matt, Democrats Like To Mix Politics & Social Media More Than Republicans, Independents, published September 4, 2012. https://marketingland.com/politics-social-media-pew-study-20444.
[7] Russell, A., U.S. senators on Twitter: asymmetric party rhetoric in 140 characters, American Politics Research, 2017.
[8] Shapiro, M. A., Hemphill, L., Politicians and the policy agenda: does use of Twitter by the U.S. congress direct New York Times content?, Policy Studies Organization, 2016.
[9] Word2Vec, Doc2Vec & GloVe: Neural Word Embeddings for Natural Language Processing, https://deeplearning4j.org/word2vec.html, published 2017.
[10] Vote Smart Application Programming Interface, https://votesmart.org/share/api#.Wpm1TOjwY2w, published 2018.