DSCC 465 Kaggle Competition Report

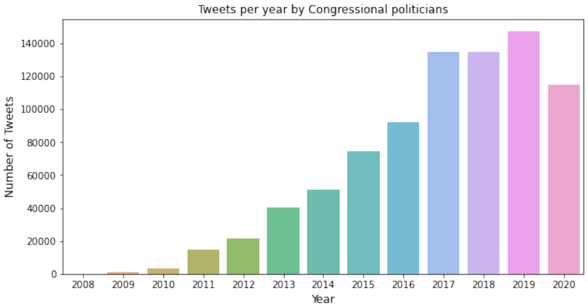
Lisa Pink and Miguel Novo Villar

***Abstract*—The following work outlines an analysis of our team's process, performance, and evaluation in the Congressional Tweets Kaggle Competition.The researchers used natural language processing techniques to clean the data and then fit machine learning models to predict tweets belonging to either the Democrat party or the Republican party. There was also a comparison in the performance of different algorithms in terms of time and accuracy of predictions.**

I. INTRODUCTION

The aim of this project was to classify congressional tweets based on the political affiliation of the Twitter user. The dataset that was used in this study was extracted from Twitter and consists of every tweet from Congressional politicians on Twitter between the years 2008-2021. The entire dataset contains 857,803 tweets, which was split into a training set of 592,803 tweets and a test set of 265,000 tweets.

Fig. 1. Total Congressional Tweetds Per Year



The test set contains 312,116 tweets from Democrats and 261,975 from Republicans. The most common hashtag in the entire training dataset is "COVID-19" with a total repetition of 16,717 times.

The six attributes in the dataset are ID, favorite count, full text, hashtags, retweet count, and year. These features help predict the class label, which is the party of the owner of the tweet, classified as "D" for Democrat and "R" for Republican. For our analysis, we only used the text feature, as the other features would not aid in predicting party.

Using a combination of natural language processing techniques to clean the text of the data, and testing a series of machine learning models to fit the cleaned data, we were able to predict the political party of the congressional tweet in the test set with an accuracy of 0.84712.

II. DESCRIPTIVE ANALYSIS

Social media has changed the way we communicate and politics are based on communication and conveying. Consequently, since the elections of 2008, politicians have adopted new methods of reaching out to the audience in the digital age. Bimber (2014) points out how Barack Obama made the most sophisticated and intensive use of new mediums of communications and analytics in his campaign; which later shaped the future of digital media campaigns.

The increased use of social media apps such as Twitter and Facebook has also contributed to creating an ever-growing divide between the Democrat and Republican parties in the United States. Today, more than ever in the past, citizens are easily able to express their political opinions at the touch of a button, and have those opinions reach the masses. As part of our initial analysis, a word cloud was created for each political party. The word clouds are presented as follows:

Fig. 2. Democrat Word Cloud

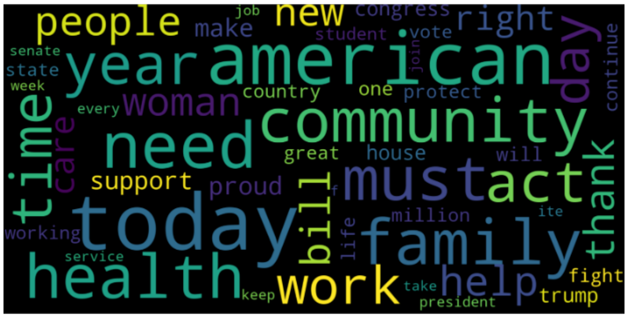
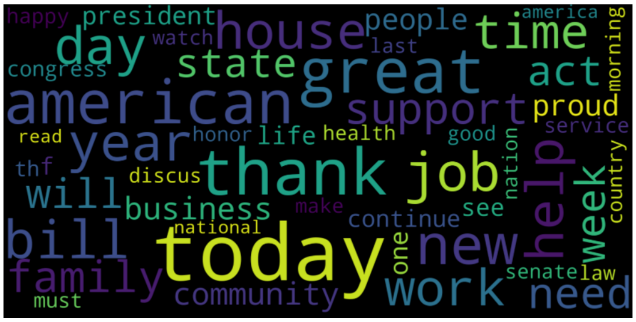


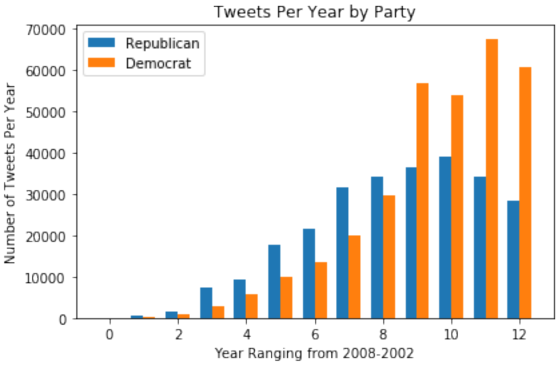
Fig. 3. Republican Word Cloud



The difference in word composition between the two word clouds reflects the differing attitudes and values between the two parties. The Democrat tweets contained many repetitions of words such as "community", "health", and "woman", while the Republican tweets repeated words such as "great", "job", and "business". These words are consistent with the traditional values of the two parties.

After observing the difference in the types of words tweeted by each party, the analysis shifted to focus on the volume of tweets, favorites, and retweets from each party over the course of the years in study. As we can see in the following graphs from our dataset, since 2008 the total number of tweets per year by congressional politicians, as well as the average number of retweets, has significantly increased and therefore matches our previously stated assumptions regarding digital media campaigns. However, the volume of tweets changed at different rates for the two parties.

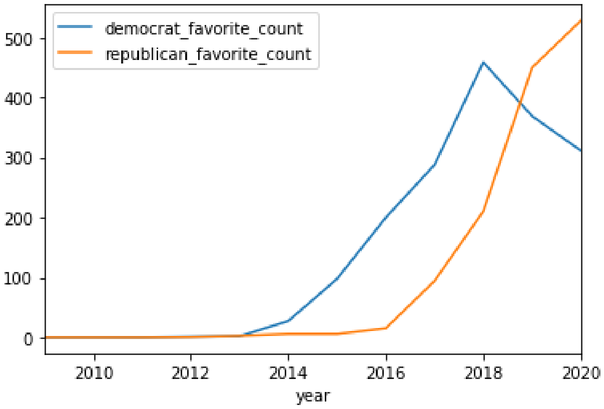
Fig. 4. Tweets Per Year by Party



The number of tweets per year increased rather exponentially for Democrats, while the Republican data saw a far less dramatic and more leveled-off increase. Recent years even saw a decrease in tweet volume from Republicans while Democrats continued to increase/level off.

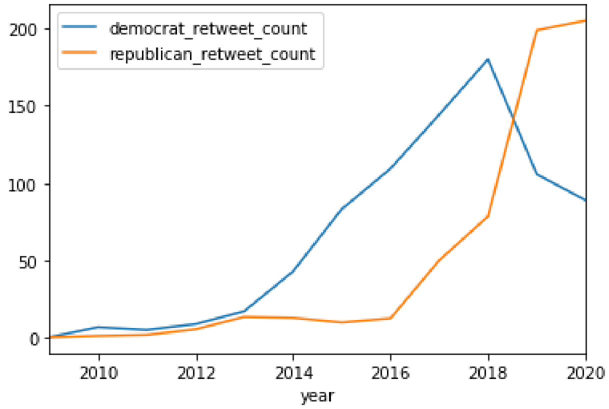
The next part of the analysis looks at tweet performance and popularity between parties. As the use of Twitter became more popular, the tweets began to gain traction in terms of favorite count.

Fig. 5. Favorite Count Times Series



Democrat congressional tweets began to gain significant number of favorites around 2013, while Republican congressional tweets saw a dramatic increase in favorite count in 2016, which is the year that Donald Trump was elected president, representing the Republican party. Since 2016, the favorite count continued to increase for Republicans, while Democrat favorite count spiked around 2018 and has been in decline since then. As can be seen in the following image, retweet count follows this exact trend and the two figures are nearly identical.

Fig. 6. Retweet Count Times Series



III. CLASSIFICATION METHODS

The first step in our classification was preparing the training data by cleaning the text of the full text feature. We began by using a function to remove stop-words, contractions, mentions, the character “rt” and “RT”, entity references, informal abbreviations, emoji’s, links, alphanumeric, words shorter than two letters, white space and quotation columns from every single string. Stop-words and other non-meaningful text don’t provide information and can be removed without affecting the meaning of the sentence from a computational perspective. Removing these words makes the remaining text less noisy and more meaningful. At this point, our text was clean and ready to be tokenized, POS-tagged, and lemmatized.

Tokenizing text is to break sentences into words, POS-tagging is tagging each word according to its part of speech, and lemmatization is reducing words to a normal form. These three methods, when performed on the cleaned text, aid in finding patterns and conducting the language processing.

Having descriptive and meaningful strings of words we performed TF-IDF, which stands for “Term Frequency and Inverse Document Frequency” which reflects the relevance of a string in a document. After having a vectorized version (therefore a matrix form) that can be interpreted by algorithms we performed different algorithms like support vector machines, k-nearest neighbors, neural networks and logistic regression to binary classify each tweet. The highest accuracy according to our testing dataset against Kaggle’s predictions came from a logistic regression model with an accuracy of 0.84712.

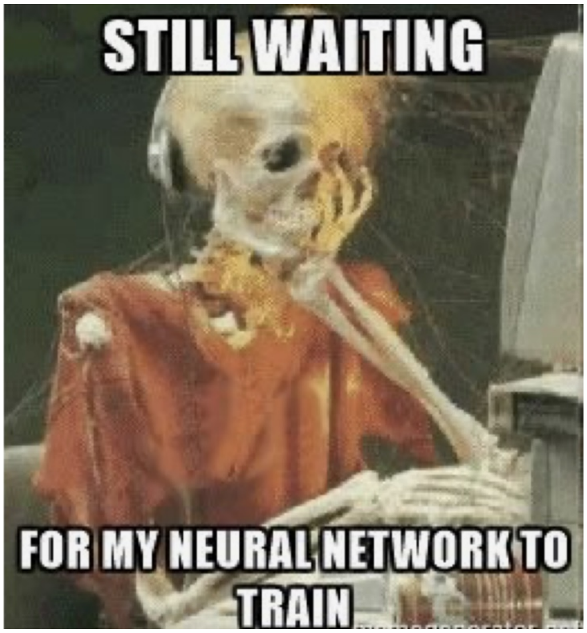
The models that were created in this study were as follows:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- SVM
- Naive Bayes
- Ensemble methods

Our top performing models were the logistic regression model with the newly preprocessed data, and two ensemble methods. The ensemble methods included SVM and logistic regression models, and the predictions were determined by the voting of the models. The random forest model was created but was not implemented as the time to run was too long. Decision tree models had lower accuracy than other models.

IV. CLASSIFICATION RESULTS

Over the course of the two weeks working on this project, many different types of classification models were built and tested. We also had a few neural networks built, but those ended up taking too long to train, and thus were not used in the end. The first few dozen initial models seemed to level out at around .79 accuracy. We tried various types of models and continued to remain stuck around the same accuracy. At this point, we ruled out the possibility that our issue was the modeling component, and took a closer look at the preprocessing of the data. In order to improve our accuracy, we added components to the data cleaning and preprocessing process and tried different vectorizers. We then ran another logistic regression model through the newly cleaned and vectorized data, which increased the accuracy to 0.84712. We would have liked to run more models on the new data, but the time for the competition submissions came to a close.



V. CONCLUSIONS AND LIMITATION

Among many different improvements and techniques that could have been approached, the researchers identified neural networks as a technique that could have potentially improved the results. From our trials on the training dataset, the researchers were able to obtain an accuracy of 0.75 on the testing dataset with simply 20.000 samples from the training dataset. Unfortunately, training the deep learning model on the entire dataset was not feasible.

Reflecting on our results, a classification that yields a .847 is not an accurate prediction and there is a large room for improvement. Nevertheless, using the variance of the stochastic average gradient descent solver we were able to accomplish a rapid prediction that could justify the variance tradeoff. The biggest limitations in this study were the time and computing resource constrains. The majority of the models took at least an hour to run, and more complicated models, such as neural networks and ensemble methods, were not able to run in time before the end of the competition.

REFERENCES

[1] Bimber, Bruce. "Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment." Journal of Information Technology Politics 11.2 (2014): 130-50. Web.

[2] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T.E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

[3] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit- learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

[4] Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

[5] Manna, S., Nakai, H. (2020). Comparative analysis of different classifiers on crisis-related tweets: an elaborate study. In Nature-Inspired Computation in Data Mining and Machine Learning (pp. 77-94). Springer, Cham.

[6] Aborisade, O., Anwar, M. (2018, July). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 269-276). IEEE.

[7] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. Plos one, 16(2), e0245909.