Congressional Tweet Classification

Derek Caramella[1] and Stefano Bastianelli[2]

***Abstract*— We extract, transform, and analyze over 857,000 records to classify a tweet's owner as a Democrat or Republican. We utilized the Logistic regression technique that exhibited 88.884 percent accuracy. We conclude that a tweet's content can reveal the owner as Democrat or Republican.**

I. INTRODUCTION

Twitter is a social media platform heavily utilized by the United States & Japan; it offers user microblogging & social networking to interact with the community [1]. Influencers, actors, politicians, & average individuals congregate on the application to post their thoughts on current events & interests. We hypothesize that the user's associated political party can be identified by a tweet contents.

We utilized NLP techniques, such as Non-Negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), & Term Frequency–Inverse Document Frequency (TF-IDF) to classify the tweets.

After processing over 857,000 tweets, we identified five distinct tweet topics: (1) health care, (2) government, (3) gratitude, (4) business, & (5) community. We uncovered that the majority of tweets contain the words: small business, year, health care, thank you, America, & house. Moreover, we recognized that tweet length & frequency gradually increased year-over-year. Lastly, our data set contained approximately equivalent class representation: 55% of tweets originating from Democrats & 45% posted by Republicans.

II. DATA

A. Descriptive Statistics

First, let's consult the Twitter data-set by exhibiting the descriptive statistics:

Party	Metric	Tweet Length	Retweet Count
Democrat	Min	0.000	0.000
	Mean	96.692	111.331
	Median	86.000	6.000
	Max	282.000	3,315,178.000
	Std.	43.740	4,455.090
Republican	Min	0.000	0.000
	Mean	83.589	74.865
	Median	74.000	4.000
	Max	248.000	1,166,71.000
	Std.	40.547	703.736

TABLE I
DESCRIPTIVE STATISTICS

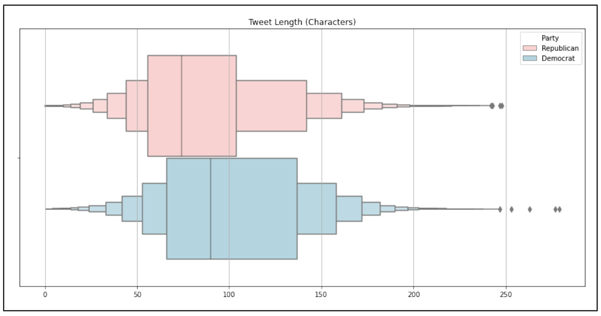

Fig. 1. Tweet Length Boxen Plot

From the Descriptive Statistics, the average tweet length is roughly equivalent between Democrats and Republicans. Moreover, for both parties, the tweet length is right (positively) skewed. Lastly, the standard deviation of tweet length is nearly congruent between parties. The Democrat's tweets receive, on average, more retweets compared to the Republican's. However, the Democrat's retweet standard deviation nearly quadruples the Republican's standard deviation. The data-set is saturated by the Democrat (55%) relative to the Republican class (45%). From the descriptive statistics, we conclude that the Democratic tweets are longer and receive more interaction relative to the Republican tweets.

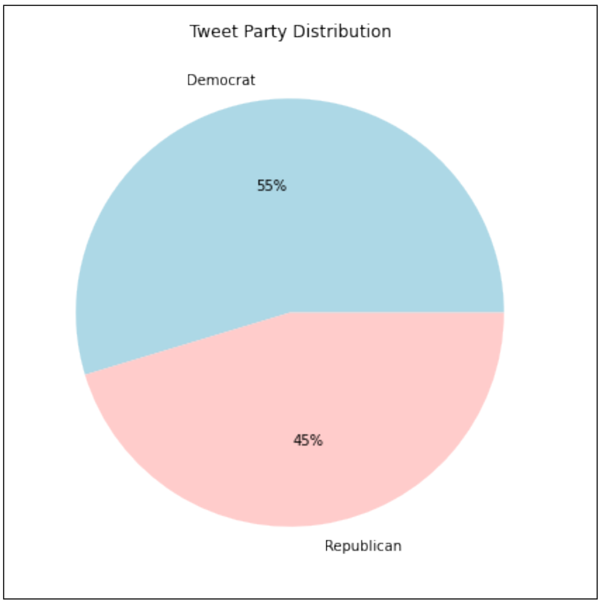

Fig. 2. Class Imbalance

Of all tweets, the most used words/phrases are small business, thank you, look forward, year, health care, stop, & America. Although the word cloud provides surface-

*This work was supported by the University of Rochester

[1] D. Caramella is a Functional Analyst with Tiber Creek Consulting University of Rochester
derekcaramella@gmail.com

[2] S. Bastianelli is the IT Systems Manager for the Residential Life Department at the University of Rochester
stefano.bastianelli@gmail.com

level insight about the tweets, LDA & NMF provides topic analysis to uncover the general ideas behind the tweets.

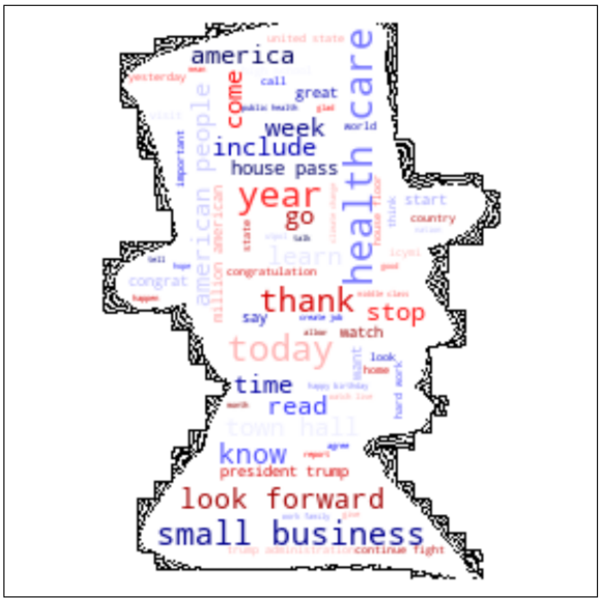

Fig. 3. Tweet Word Cloud

B. Latent Dirichlet Allocation (LDA)

LDA is an unsupervised learning topic modelling algorithm used in Natural Language Processing (NLP). LDA approaches text as topics relative to word counts, LDA reduces dimensionality by compressing a corpus to topic$_i$: Weight(topic$_i$, T). LDA utilizes the Dirichelt prior on top of the data generating process. Following LDA, t-SNE can visualize the data by preserving local structure by minimizing the Kullback-Leibler divergence (KL divergence) [4]. We utilized the Count Vectorizer to convert the review text to vectors, we ignore terms with a higher frequency of 95 percent, removed words with a frequency count less than 2, and posited 5 disparate clusters. We implemented a learning offset of 50 and set 10 maximum iterations for the LDA model. The top 10 words in each topic is depicted below.

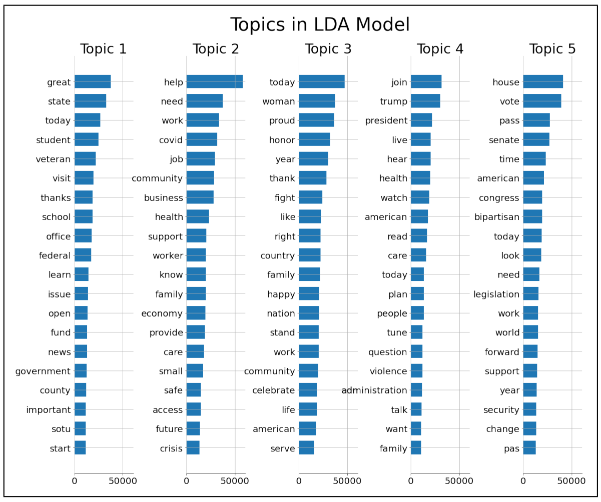

Fig. 4. Review LDA 5 topics.

We labeled the topics as (1) health care, (2) government, (3) gratitude, (4) business, & (5) community.

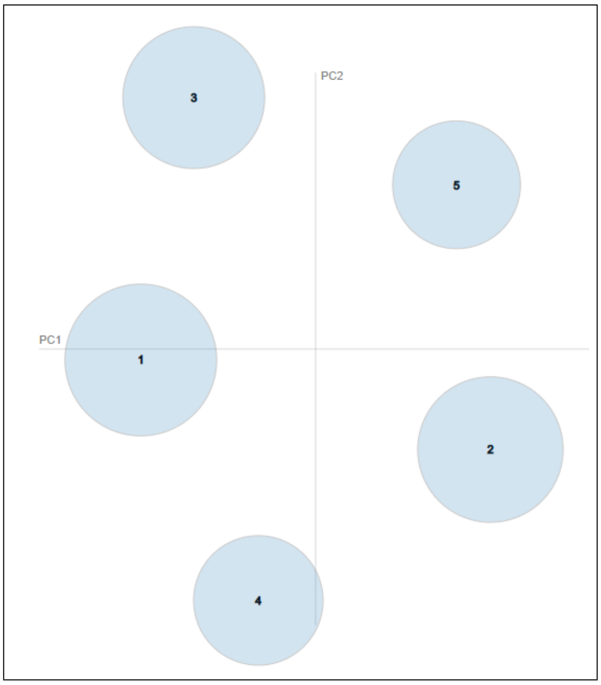

Fig. 5. LDA TSNE

C. Non-negative Matrix Factorization (NMF)

NMF is an unsupervised topic modeling algorithm. A matrix V is factorized into two matrices W and H, such that three matrices contain no negative elements. By factoring the matrices, the factor matrices' dimensions may be significantly lower than V. We implemented the Frobenius norm and the Kullback-Leibler beta divergence [3]. The top 10 words are exhibited below.

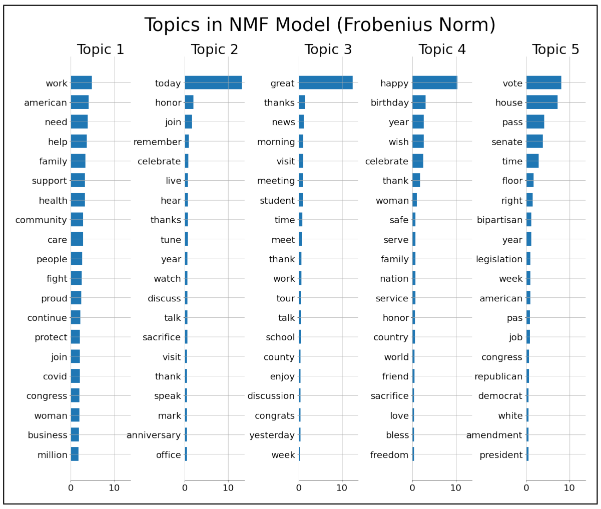

Fig. 6. Review NMF 5 topics.

We labeled the topics as (1) government, (2) veterans, (3) government, (4) gratitude, & (5) celebration.

D. *Time-Series Analysis*

We observed a significant overall increase in Twitter activity from 2008 to 2020. Tweet length & post frequency increased over the time domain. The highest growth occurred from 2015 to 2016. The Democrats overtook the Republican's in tweet frequency in 2016; however, prior to 2016, the tweet frequency was roughly equivalent. Following 2016, the Republican's held roughly consistent post counts.

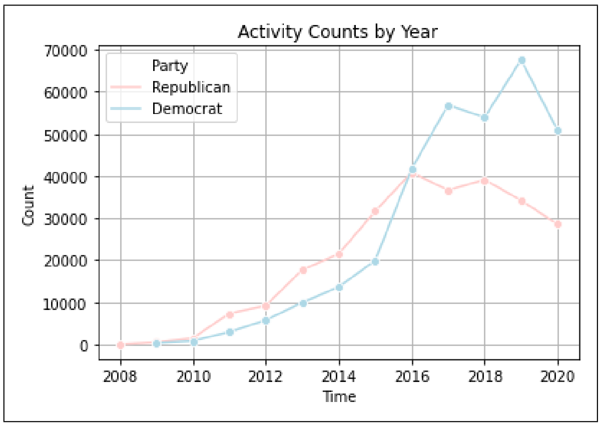

Fig. 7. Tweet Activity

The tweet length was roughly equivalent between the Republicans & Democrats. On average, the length of tweet's increased, the largest increase occurred from 2017 to 2018.

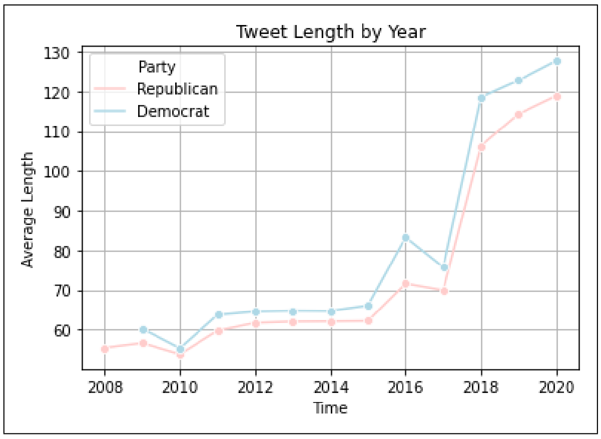

Fig. 8. Tweet Length

III. METHODS

We employed the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm to extract features from the retweet corpuses. TF-IDF applies weights to the frequency a term appears in a document; however, the result is offset by the number of times the term appears in the entire document – subsequently, stop words are punished.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

where $tf_{\mathrm{x,y}}$ is frequency of x and y, df_{x} is the number of documents containing x, and N is the total number of documents. Prior to model training, we lemmatized each tweet. We removed special characters, words less than three characters, emojis, stopwords, & replaced contractions with the expanded statements. Additionally, we stripped whitespace from the sentences, lastly we replaced markdown text, such as "/amp" with "and". We passed each tweet into the TF-IDF object, n-gram=(1,2) or n-gram=(1,1) & maximum features=30,500. Lastly, we analyzed the concatenated attribute utilizing `Vader`, which produces a sentiment strength based on the inputted text listing negative, neutral, positive, & compound sentiments [8]. We converted the sentiment dictionary to a 1xn array and concatenated the TD-IDF to the sentiment array.

Following tweet pre-processing & feature engineering, we passed the concatenated arrays through a variety of models listed in Section IV. RESULTS. We conducted a grid search on the logistic regression & uncovered the best parameters were an elasticnet, with an l1-ratio of 0.5. Moreover, we experimented with a Tensor Flow sequential model with various depths and widths; however, our results were overfitting; consequently, producing a non-generalizable model [5]. Lastly, we attempted to create a random forest with over 20,000 features, but the results did not outperform the regularized logistic regression [6]. We discuss our results further in Section IV.

IV. RESULTS

We constructed a logistic regression model with 88.884 percent accuracy that classifies a tweet's owner as a Democrat or Republican.

We successfully constructed disparate topic clusters: (1) health care, (2) government, (3) gratitude, (4) business, & (5) community. We built a machine learning pipeline that lemmatizes, filters, scales, & extracts features from the tweets that classifies the owner's political party affiliation.

Word Vectorizer	Model	Accuracy
TF-IDF (n-gram: 2)	Logistic Regression	85.259
TF-IDF (n-gram: 1)	Logistic Regression	85.045
Count Vectorizer (n-gram: 1)	Logistic Regression	88.884
Count Vectorizer (n-gram: 2)	Logistic Regression	85.045
BERT (30 epochs) [8]	BERT (30 epochs) [8]	75.560
TF-IDF (n-gram: 2)	Sequential (5 Layers, 256 Nodes)	85.338
TF-IDF (n-gram: 2)	Random Forest	87.328
Train N: 592,803 Test N: 265,000		

Twitter enables users to comment on events & that provides users the ability to interact with the community. We utilized NLP techniques & the logistic regression to classify a tweet's owner as a Democrat or Republican with 88.884 percent accuracy.

References

[1] Cao, Q., Duan, W. and Gan, Q., 2011. Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. Decision Support Systems, 50(2), pp.511-521.

[2] Daniel D. Lee, H. Sebastian Seung, 2001. Algorithms for Non-negative Matrix Factorization.

[3] Patrik O. Hoyer, 2004. Non-negative Matrix Factorization with Sparseness Constraints. Journal of Machine Learning Research 5, pp.1457–1469

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan, 2003. Latent Dirichlet Allocation.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[6] Hui Zou, Trevor Hastie, 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 67 (2), pp.301–320

[7] Anita Kumari Singha, Mogalla Shashi, 2019. Vectorization of Text Documents for Identifying Unifiable News Articles. International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019

[8] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014