Congressional Tweet Partisnship Classification

Jiahang Wu and Qihao Yun

I. Introduction

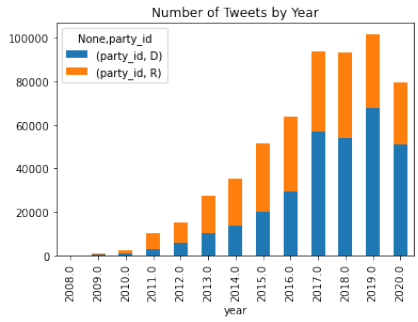
This report analyzes and explores a US twitter dataset from April to December 2020, focusing on Covid-related analysis. The report hypothesizes twitter discussion of covid related topics are connected to geographical differences, as it is the case with the US political differences. The report finds substantial evidence against this hypothesis and discusses possible alternativer insights from the dataset.

II. Descriptive Analysis

The Twitter sample contains 200,000 twitter posts during 2020. This report randomly selects 50,000 samples and explores the dataset. Only a small part of the features directly describes the tweet post; many others are information about the poster which is convenient for this report to explore the tweets.

The dataset of this study is a collection of twitter tweet texts and some accompanying information: favorite count, hashtags, retweet count, year, and party affiliation. The training dataset contains 592,803 rows while the testing dataset 265,000. For the purpose of predicting partisanship, we observe some imbalance within the data. First, the number of data concentrates substantially in the 2014-2020 range; consequently, models trained over the entire dataset would likely favors tweets from later dates. This is somewhat unsurprising as twitter – along with the internet industry in the US – experienced a user boom up until very recently.

We also notice that there are more republican tweets up until 2017 when Democrat tweets overtook by a large margin consistently. We suspect this may be related to the polarizing 2016 US election.



	Democrat	Republican
2008	nan	13.0
2009	333.0	511.0
2010	816.0	1461.0
2011	2935.0	7325.0
2012	5731.0	9201.0
2013	9937.0	17674.0
2014	13634.0	21454.0
2015	19877.0	31613.0
2016	29581.0	34205.0
2017	56882.0	36637.0
2018	53914.0	39126.0
2019	67590.0	34214.0
2020	50886.0	28541.0

Fig. 1. Number of Tweets Grouped by Year and Partisanship

In our naïve regression, we find that hashtags within the texts contain significant information in identifying partisanship. We find that some of the most frequent hashtags are inherently partisan, such as "#tcot" – "top conservatives of twitter". We also find that 75% of the hashtags have less than 3 mentions, and only the top 6.25% hashtags have more than 16 mentions in the training dataset. The top 10 hashtags are political and constituted as just below 90% of all hashtag frequencies in this dataset (this dataset has no tweet without hashtags).

The most frequent hashtag is "covid19" which is a natural result of the pandemic since 2020. The findings suggests that there is substantial skew in the hashtag data for tweets of 2020, the last year of this dataset.

The findings suggest that it may be appropriate to separate the dataset into two or more sets to improve performance for under-represented sample.

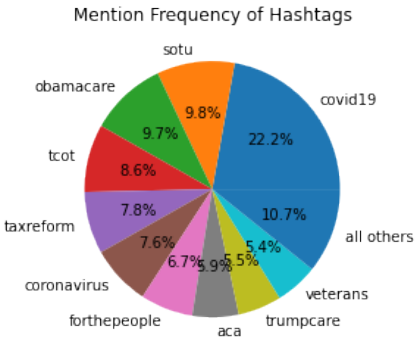


Fig. 2. Document Frequency (Percentage) of Top Hashtags

TABLE I
Document Frequency of Top 10 Hashtags

covid19	16954
sotu (State of the Union)	7524
obamacare	7449
tcot (Top Conservatives on Twitter)	6553
taxreform	5987
coronavirus	5830
forthepeople	5092
aca (Affordable Care Act)	4475
trumpcare	4206
veterans	4144

Other features of the dataset have near-zero correlation with party identity, and they are ignored in this study as a result. PCA as a summarization tool is unavailable in this study due to the large size (>300 GB) of the vectorized text of hashtag in dense matrix format.

TABLE II
CORRELATION BETWEEN PARTISANSHIP AND OTHER FEATURES

	Favorite Count	Retweet Count	Year	Party ID
Favorite Count	1.00	0.756996	0.059132	-0.023679
Retweet Count	0.756996	1.00	0.045455	-0.019159
Year	.059132	0.045455	1.000000	-0.210004
Party ID	-0.023679	-0.019159	-0.210004	1.000000

III. CLASSIFICATION METHODS

A. Vectorized Text Logit Regression

1) This is a naïve, brute-force approach of 5-fold CV logit regression on party identity with the tf-idf vectorized tweet texts (n_gram = (1,2)) as the only input matrix.

We begin the process by cleaning and standardizing the tweet texts, removing punctuation, emoji, and other non-standard elements we do not plan to interpret with our methods. The text is then tagged, lemmatized, and then vectorized as a (1,2) gram Tf-idf vector.

B. Vectorized Hashtag and Sentiment Logit Regression

This is a 5-fold CV logit regression on vectorized hashtags. Additionally, we calculated the simple sentiment value (positivity and negativity) based on the tweet texts and two lists of positive and negative words. The values are appended as new columns of the input matrix to the regressor.

C. Ensemble of Two Logit Regressions

The prediction of this model is based on the average prediction probability of the two models above. This is our best performing model.

IV. CLASSIFICATION RESULTS

Surprisingly, the brute force method achieved pretty good results. However, the result of this method is directly contingent to the cleaning process of the texts. In fact, this method slightly punishes thorough cleaning.

Initially, we adopted the perspective of NLP analysis in the cleaning process and removed hashtags, misspellings and other elements. The key motivation was to suppress the growth of dimension from vectorization process: reducing training costs and mitigating other issues associated with high dimentionality thusly.

Most of the work is done through regular expression that handles cases individually. Misspelling detection is very costly, so we alternatively configured the vectorizer to ignore low document frequency tokens (min_df = 3). In variations of this preprocessing, we generally achieve 75-80% accuracy on the test dataset. This is because information –especially hashtags – are lost in the cleaning process.

Inclusion of hashtags and misspellings improved the model's accuracy to 85-88%, albeit at the cost of training time as a result of larger dimensions. This observation indicates that hashtags are substantial contributors to explaining partisanship, and this inspired our second model in lieu of a topic extractionist model. Misspelling also contributes to explaining partisan affiliation; however, we are unable to further validate this thought due to the very high cost of detecting misspelling and then finding correlation with model performance at scale.

We understand the inclusion of sentiment scores columns on the input matrix may be inadequate handling of the information, given that we are using a logit regression which treats each feature as conceptually linear contributors. Our assumption of sentiment is that they are the most meaningful when coupled with individual cases of the hashtags (or, topics) of the tweets. However, that would be intractably expensive to capture the combinations with our limited resources, and we chose to compromise with this handling.

Surprisingly, our assumption unnecessarily applies as the final coefficients associated with the sentiments remain very significant: Negativity: -1.79566403, Positivity: 0.04678373. This indicates that, at least in general, sentiment in the tweet text does correlate with partisanship independent of the topics discussed. This model yields approximately 85% accuracy with the test dataset.

Finally, we achieved our best results by comnbining the two models and forming a small ensemble. Because both models are logit regression, we were able to use the prediction probability from the models directly and calculate the average probability for the prediction. The resulting accuracy of this model is 89.7%, marginally lower than 90% and places our model at a relatively higher position among all participants. We believe this indicates that we have captured some information not learned by both models through this ensemble, despite both models do include hashtags in the input dataset in different forms.