# Text Classification of political party affiliation using Tweets

**Introduction**

Although the two major political parties in the United States are built on seemingly opposing principles of liberalism (Democrat) and conservatism (Republican), these lines have been increasingly blurred in recent times. Complex socio-political conundrums like abortion, the COVID-19 pandemic and President Trump’s election are hot topics of discussion, further amplified by social media like Twitter and Facebook. Politicians themselves have realized the benefits of pandering to their niche voter base – using their social media platforms to increase their reach amongst voters. Therefore, it is increasingly valuable to be able to classify tweet content in order to find potential leads and target audiences.

They primary stakeholders for such a classification would be companies involved in OSINT (open-source intelligence tool) like political consulting firms to gauge public sentiment. It could also be extended to classify target audiences by classifying users as Democrat or Republican. When integrated with other identifying information using cookies (location, browsing history, device, age, gender, etc.), political affiliation could be used effectively to target audiences with political ads, fake news, propaganda, and even ads for products that show a skew towards a particular party affiliation.

Several companies have already identified the value in such a prediction; in fact, Cambridge Analytica used similar models (albeit much more complex) in 2014 to harvest data and build voter profiles. Models deployed on large quantities of data can often yield actionable insight for political campaigns. The aim of this analysis is to evaluate the performance of several models on Twitter posts made by US Senators and politicians.

**Analysis and Models**

*About the Data*

The most challenging (and expensive) problem in building an accurate model is obtainining and labeling a large dataset of Tweets to train them on. Since labelling such a large number of tweets is a time-consuming and expensive process, an alternate approach was necessary to circumvent this challenge. Twitter handles of 581 politicians were scraped from three different websites. These websites also contained the political affiliation of the politicians. The TWINT API allowed the entire post history to be harvested from just the twitter handle. The true label of each of the tweets was assumed to be the political party affiliation of the politician that posted them.

Approximately one million tweets were obtained using the method stated above to construct a dataset.

Figure 1: 900,000 tweets were collected in total, and were moderately balanced between the two parties..

Table 1: Metadata collected along with tweets include details of the tweet; however not all information may necessarily be present due to user privacy settings.

|  |  |
| --- | --- |
| **Variable** | **Information** |
| **created\_at** | Time of tweet creation. [UNIX timestamp] |
| **date** |  |
| **time** |  |
| **timezone** |  |
| **username** |  |
| **name** | Registered Twitter name |
| **place** | Common name of location of tweet (if available) |
| **tweet** | Text of length < 280 characters |
| **language** | Languange of tweet |
| **mentions** | list of @mentions |
| **urls** | list of urls mentioned in tweet |
| **photos** | names of photo files attatched to tweet. |
| **replies\_count** | # of replies |
| **retweets\_count** | # of retweets |
| **likes\_count** | # of likes |
| **hashtags** | list of #hashtags in tweet |

Wordclouds can give us a qualitative view of the most commonly occurring terms. The below plots show the most frequently occurring terms in the tweets.

Text

Description automatically generated

Figure 2: Word cloud for tweets sent by Democrat politicians in the United States.

Text

Description automatically generated

Figure 3: Word cloud for tweets sent by Republican politicians in the United States

*Models*

The below three models were trained using different vector inputs:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Vectorization** | **Binary Vectors** | **TFIDF Weighted Vectors** | **Model Code** |
| Multinomial Naïve Bayes | Count Vectorizer (sklearn) | No | No | **MNB** |
| Multinomial Naïve Bayes | TFIDF Vectorizer (sklearn) | No | Yes | **MNBTF** |
| Bernoulli Naïve Bayes | Count Vectorizer (sklearn) | Yes | No | **BNB** |

All models were evaluated using 5-fold cross validation. True Positive Rate (TPR), False Positive Rate (FPR), F1-Scrore and Model Accuracy were collected from each fold, and averaged across runs to evaluate model performance.

|  |  |  |
| --- | --- | --- |
| **Model** | **F1-Score** | **Accuracy** |
| **BNB** | 0.65 | 0.63 |
| **MNB** | 0.68 | 0.63 |
| **MNBTF** | 0.71 | 0.63 |

**Results**

Of the three models, the TFIDF weighted Multinomial Naïve Bayes (MNBTF) model showed the highest F1-Score. All three models performed with relatively similar accuracy. Below are the confusion matrices for the 3 models:

Table 4: Confusion Matrix for BNB model

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | Dem | Rep |
| **True** | Dem | 335,939 | 208,600 |
| Rep | 152,905 | 273,528 |

Table 5: Confusion Matrix for MNB

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | Dem | Rep |
| **True** | Dem | 388,812 | 155,727 |
| Rep | 204,307 | 222,126 |

Table 6: Confusion Matrix for MNBTF model shows that the model tends to misclassify more republican tweets as being democrat. It is however significantly better than the other two models.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted** | |
|  |  | Dem | Rep |
| **True** | Dem | 446,936 | 97,603 |
| Rep | 262,836 | 222,126 |

Chart, scatter chart

Description automatically generated

Figure 4: MNBTF Model ROC Curve

The ROC Curve of the MNBTF Model shows that the even the best model that was trained shows a moderate performance in distinguishing between the two classes.

**Conclusions**

In this analysis, approx. one million tweets were collected through the use of an API, and labeled using their authors political party allegiance. Tweets were collected from 582 US Politicians who are relatively popular twitter users including Joe Biden, Maxine Waters, Mitt Romney and Mitch McConnell. Three models were run (all using different types of data normalization) and validated. Although the Multinomial Naïve Bayes model (with TFIDF weighting) performed the best, it was still not an accurate enough model to be put into production. In order to improve model accuracy, it would be necessary to collect a larger amount of data, crowdsource “true” labels, and evaluate other types of models. Nevertheless, the models generated can be used to make a better prediction than one made at random.

**Appendix A**

Text

Description automatically generated

Figure 5: Wordcloud of complete dataset (Democrat + Republican)

**Chart

Description automatically generated**

Figure 6: BNB Model ROC Curve.

Chart, scatter chart

Description automatically generated

Figure 7: MNB Model ROC Curve