Applied Analytics Frameworks and Methods II  
Project Deliverable 2

Gregorio Galletti, Katherine Li, Marsya Chairuna

short line

# 1 Introduction

Twitter has been the most popular social media platform for lawmakers and politicians to influentially express thoughts to the public. Meanwhile, the contents of tweets could help audiences understand what connects with constituents starting from examining the frequency of most used words.

As the 2020 election is approaching, this project is useful to understand a political agenda amongst potential presidential candidates. This project can also be beneficial for research to analyze the respective sponsors of each political party. **This project aims to (1) predict the appropriate political parties (Democrat and Republican) given published arbitrary tweets and (2) compare the sentiment differences between two parties in topics: climate changes and fossil fuel.**

# 2 Methodologies

2.1 Experimental Design

In order to perform our analysis, we conducted the following steps:

1. **Data collection:** The original dataset contains two files. ExtractedTweets.csv extracts tweets from all the representatives (latest 200 as of May 17, 2018) with the following variables: Party, Handle, and Tweets. This dataset is extremely broad and comprehensive with 84,502 tweets from 433 different representatives (50% Democratic and 50% Republican). The TwitterHandles.cvs extracts a list of Democrat and Republican Twitter handles and metadata with the following fields: Party, Name, Handle, and Avatar URL.
2. **Additional data processing:** 
   1. **Split the original dataset:** We started with splitting the data into two parties, Democrat and Republican to easily identify the differences in the lexicon and its frequency between the two parties. We also kept a data frame comprehensive of all of the tweets.
   2. **Clean data:** we converted all the letters to lowercase, removed punctuation, URLs, numbers, empty words, and white space. A corpus was made for each data frame (Republican, Democrat, and all parties).
   3. **Stemming:** to avoid redundant data with similar meaning, a dictionary for each subset of tweets.
   4. **Convert object**: Data frames were converted into Document Term Matrixes.
   5. **Remove sparse terms**: We eliminated works with low frequency by setting a sparse value of 0.975, which means that terms appeared less than 2.5% of the dataset were removed.
   6. **Other**: (i) For all\_tweets dataset, we removed the handle, only remaining predictors and labels for the prediction. (ii) For classification models, the term “trump” was removed as we found that it might not be the most insightful primary term determining Democrat and Republican in tree models.
   7. **Split the dataset:** For the series of classification models, we split the dataset 70% into the training set and 20% to the testing set randomly using *sample ( ).*

2.2 Techniques Selection

1. **Classification model**

To predict political parties (Democrat and Republican), we used and compared the performance of three supervised machine learning techniques: decision tree, random forest, and logistics regression. This will also give opportunities to interpret results from different perspectives.

* 1. **Logistics Regression**: Logistics regression is commonly used for classification problems. Although it is one of the simplest in the classification technique family, we include a logistic regression model as a practical baseline into the research while using other complex classification models and it provides insights about how well complex models perform beyond the simplest baseline.
  2. **Decision Tree**: We chose decision trees because it is particularly useful when there are a large number of independent variables and when there are likely to be non-linear relationships amongst variables as logistics regression may be difficult to use in this situation. Additionally, trees are better at interpreting results than regression as discovering the rationales that distinguish Democrats and Republicans are essentially important in our study.
  3. **Random Forest:** As trees generally do not have the same level of predictive accuracy as other classification approaches, random forest, a way of aggregating many decision trees with random selection, will help to improve the accuracy of results. Second, more trees also mitigate the situation of overfitting trees in the model. Last but not the least, random forest has higher power to handle a larger dataset with wide dimensionality, which is suitable for this case when there are more than 40 independent variables.

1. **Sentiment analysis**

Sentiment analysis is suitable for investigating the Democrats and Republicans’ sentiments and options toward the topics “climate changes” and “fossil fuel” on Twitter as the technique is useful in social media monitoring. In political practices, sentiment analysis has credibility and practicability. For instance, the Obama administration utilized sentiment analysis to capture and measure public opinion to policy announcements, political discussion to certain topics, and campaign messages at the 2012 presidential campaign.

Thus, sentiment analysis is an essential part of this research as it allows us to compare the sentiment differences between Democrats and Republicans on certain topics, given the audience's strong psychologically insightful understanding ahead of the 2020 presidential elections.

1. **Principal Component Analysis (PCA):**  We were working with a large set of variables ( 39 independent variables in all\_tweets data set). There are chances of finding correlations among a predictor or a set of predictors, *multicollinearity*. Therefore, to avoid any erroneous conclusion about the relevance of a predictor, we included principal component analysis to summarize the set with a smaller number of representative variables that collectively explain most of the variability in the terms used by Democrats and Republican representatives.

To build a progressive analytical research experience, we performed four models using a decision tree, random forest, and logistics regression. Then we added PCA to the four models, creating a set of another four models to see any improvement made due to PCA.

2.3 Application of Methods

2.3.1 Classification models

To predict the appropriate political parties given published arbitrary tweets, we aimed to develop a classification model for Twitter dataset. We trained and compared eight prediction models below. For all the candidate models, we observe the accuracy of the prediction on the train and test dataset. We will then observe which one is the best-performing model based on these performance parameters.

* Model 1: Decision tree model without cross validation
* Model 2: Decision tree model with 5-fold cross validation
* Model 3: Random forest classification model with 5-fold cross validation
* Model 4: Logistic regression
* Model 5: Decision tree without cross validation, with components from PCA
* Model 6: Decision tree with 5-fold cross validation, with components from PCA
* Model 7: Random forest classification model with 5-fold cross validation, with components from PCA
* Model 8: Logistic regression with components from PCA

To train the model, we split the dataset into train and test data with 70-30 composition.

**Classification Decision Tree**

Firstly, we ran a decision tree model, using rpart package, without cross validation, with a complexity parameter (cp) value of 0.0075. We named this model as Model 1. We set a grid search (cp = 0.1, 0.001, 0.001) for cross-validation of 5-fold that averages out the sample error, generating a more precise estimate. The result gave us the optimal cp of 0.001. All other parameters are set to default. We reran the decision tree model using the optimal cp and we named this model as Model 2.

**Random Forest Classification**

We then ran a random forest classification model with a 5-fold cross validation to see which mtry value, or number of variables available for splitting at each tree node, is the optimal. The resulting mtry value from cross validation is 3. We then use this mtry value and ntree value, or number of trees, equal to 300. All other parameters are set to default. We named this model as Model 3.

**Logistic Regression**

We then ran a logistic regression model using glm() function in R with a family type of ‘binomial’. All other parameters are set to default. We named this model as Model 4.

**Principal Component Analysis (PCA)**

The principle component analysis was not applied for the first four classification models. Based on the resulting scree plot, we decided to reduce our initial 55 dimensions to 10 dimensions. Next, we reran the first four models using these new components from PCA, which we named trainComponents and testComponents. We first ran two decision tree models, with and without 5-fold cross validation, which we named Model 5 and Model 8. Increasing the advancement of the tree classification model, we created a random forest from the *RadomForest* package, with 5-fold cross-validation and the number of trees (ntree) of 1000, denoted as Model 7. Lastly, we built a logistics regression model, denoted Model 8.

To avoid any possible correlation between independent variables, we conducted principal component analysis and applied its results into each of the four models described above while keeping other parameters constant. Now, we have created another four classification modes with predictor variables reduced.

2.3.2 Sentiment analysis

To study the sentiments from Democrat and Republican, we filtered all tweets about climate change and global warming for all three datasets, all\_tweets, tweets\_Democrat, and tweets\_Repliblican parties. Then we tokenized words of tweets in each dataset.

We classified the words to gain an in-depth understanding of the tweets for each party. So we used the *unnest\_tokens* function from the *tidytext* package to tokenize the tweets and a series of *dplyr* functions (group\_by, ungroup, and count, etc.) to analyze the sentiments of tweets for each party.

We first used binary sentiment Lexicon, **bing lexicon** to categorize words into positive and negative groups. However, words may reflect more than just valence as they are sentimental in different degrees. To better observe the magnitude of the sentiments for each party on the topic “climate change”, we used sentiment score Lexicon, **afinn lexicon** that assigns words with sentiment score from -5 to 5. We started with identifying the top and bottom 10 scores for each representative.

Furthermore, we used the **NRC emotion lexicon** which categorizes words based on the emotions conveyed whereas a word may reflect multiple emotions. Having a list of emotions reflected from the words used from Democrats and Republicans would help us more directly perceive the sentiments of both parties about climate change.

We then performed the same process for tweets about Fossil Fuels (filtering for “fossil fuels” ,“oil”, “natural gas”, “petroleum” and “ff”). The reasoning behind this was to identify two related controversial topics that spring juxtaposing feelings. Our hypothesis was that Democratic and Republican representatives have very different opinions about the topic: Democrats fear climate change and the negative impact fossil fuels have on it, while Republicans underplay the effects of global warming and praise the economic impact of oil and gas.

# 3 Results

# 3.1 Classification Model

**Weight of Term Frequency**

The top ten words or terms with the most weight of term frequency for Democrats are today, will, work, thank, trump, join, student, american, famili, and congress. Meanwhile, the top ten words or terms with the most weight of term frequency for Republicans are today, tax, thank, will, great, hous, bill, work, american, discuss, and act. As we can observe, some of the top words of these two parties are intersecting. The top words that are exclusive to Democrats are **trump, join, student, famili, and congress**, while the top words that are exclusive to Republicans are **tax, great, hous, bill, discuss, and act**.

Based on the weight of term frequency, tax, bill, and work have a heavy diagnostic value thus indicating importance for Republican’s tweets. On the other hand, trump, student, family, and congress have a heavy diagnostic value for Democrats.

**Results of PCA**

We used PCA to reduce the 57 dimensions to 10 dimensions. Based on the results in Model 1, Dimension 3, 4, and 5 acted as splitting nodes that differentiate Democrat and Republican. The terms that have the most contribution to Dimension 3, 4, and 5 are as follows.

* Dimension 3: last week, year, tax, must
* Dimension 4: job, tax, great, day, happi
* Dimension 5: hear, student, live, school, now

**Results of the Classification Models**

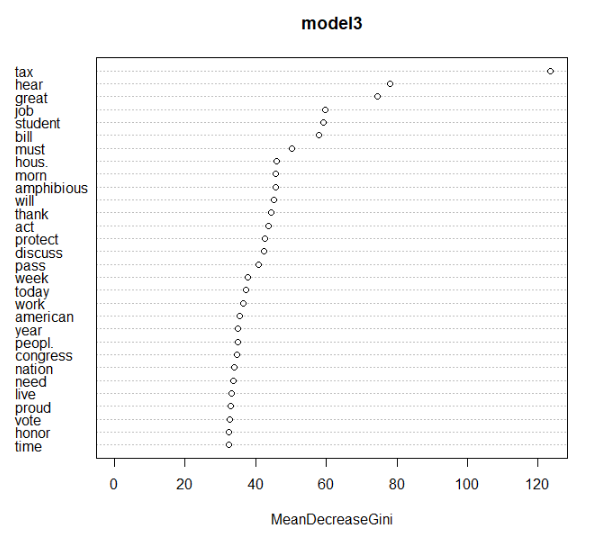
The summary of the model performances is as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Description** | **Train Accuracy** | **Test Accuracy** |
| Model1 (Decision Tree) | Cp = 0.0075 | 0.5403 | 0.5405 |
| Model2 (Decision Tree) | Cp = 0.0001, 5-fold CV | 0.5798 | 0.5796 |
| Model3 (Random Forest) | Ntree = 300, mtry = 3, 5-fold CV | 0.5910 | 0.5934 |
| Model4 (Logistic Regression) | Default Parameters | 0.5850 | 0.5827 |
| Model5 (Decision Tree) | With PCA, Cp = 0.0075 | 0.5622 | 0.5604 |
| Model6 (Decision Tree) | With PCA, Cp = 0.0001, 5-fold CV | 0.5812 | 0.5713 |
| Model7 (Random Forest) | With PCA, ntree = 1000, mtry = 2, 5-fold CV | 0.5774 | 0.5733 |
| Model8 (Logistic Regression) | With PCA, Default Parameters | 0.5586 | 0.55813 |

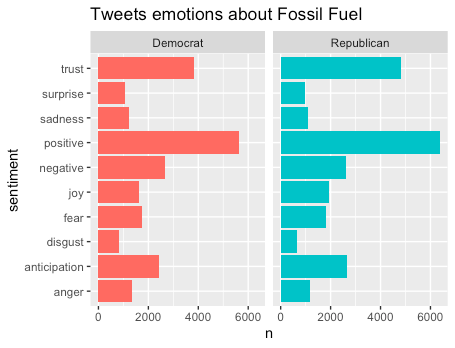
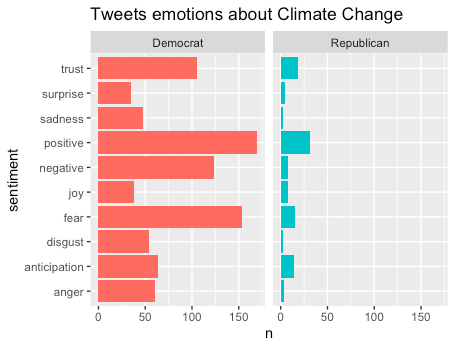
As we can observe from the table above, Model 3, Random Forest, has the best accuracy for both train and test dataset compared to all models. However, the differences between accuracy results for all models are marginal, there is no single model that greatly outperforms other models. With an accuracy of only around 60% for binary labels, our models do not perform well enough to appropriately determine which tweets come to which political party. Our model is yet to be improved in accuracy, which can be achieved by refining or adding more predictors by variable transformation or integrating with new sources of dataset such as articles and news.

**Importance Plot of the Variables**

The diagram below illustrates variable importance based on the best-performing Model 3 with Random Forest classifier. We can observe that the word “tax” has the most impact on explaining a particular observation fed to the model. This can be interpreted as “tax” has the most impact on determining the political parties based on arbitrary tweets dataset. The top six most important words are tax, hear, great, job, student, and bill.

****

3.2 Sentiment Analysis



From the two charts above it is blatant that our hypothesis was incorrect, as Republicans and Democrats seem to express similar emotional patterns about both fossil fuel and climate change. Analysis using the bing and afinn lexicons yielded similar results. In fact the average score for democrats about fossil fuels was 0.63 versus a 0.94 for republicans. Climate change did seem to separate the two parties a little more according to afinn, with republican tweets scoring on average 1.52, while democratic representatives seemed on average neutral about the subject (score = 0.1). Based on our hypothesis we were hoping to see negative scores for democrats and positive for republicans.

On the other hand, we did notice a substantial difference in tweeting patterns between the two parties. In fact, while the representatives from both parties are equally active on the social network discussing topics related to fossil fuels, Republicans are disproportionately less likely to address climate change, making up only 13% of the total tweets about the topic.

4 Conclusion

4.1 Recommendation

Our recommendation is directed towards the Democratic party, as we believe that they should better leverage Twitter to promote their ideology as well as make sure that it’s cohesive across their representatives. In this year’s Democrative primaries two recurring topics took the stage: Healthcare for all, and the need for a Green New Deal. Needless to say that, even though this dataset was a little outdated, it is alarming that the top 50 Democratic representatives lack this same type of cohesive sentiment about Climate Change and their Twitter presence seems to be no different than their opposing party.

More people are reportedly utilizing Twitter as their main source of news and information and, even though each representative has freedom of speech, it is important that Democrats establish a strong, clear and cohesive presence on the social network to transmit their beliefs and maximize their chances to capture swing voters.

4.2 Improvement and Future Research

Several improvements could be considered for future political party sentiment research.

1. The **dataset** could incorporate more variety for future research. For example, we could scrap and use more updated tweets and comments on Twitter as the trendy political topics may vary over time, which could provide us new insights on the terms that classify the parties.
2. Other social platforms could be considered as **sources of data**, such as Raddit and Facebook that accommodate reliable articles, comments, and opinions advocated from political representatives to see any significant terms that distinguish the two political parties. It will also be interesting to see how it differs from the determining linguistic terms we captured from Twitter from this research.
3. The sentiment analysis of more **noteworthy topics**, affairs, and issues addressed in politics could be conducted to add more dimension to understanding how attitudes of the two parties differ. Notably, some other trendy topics in political debates are gun control, equal pay, drug price regulation, immigration, minimal wages, etc.

# Reference

Wertz, J. (2019, January 18). Why Sentiment Analysis Could Be Your Best Kept Marketing Secret. Retrieved from https://www.forbes.com/sites/jiawertz/2018/11/30/why-sentiment-analysis-could-be-your-best-kept-marketing-secret/#2850fb982bbe

# 

# 