Project Deliverables 2 Presentation

Applied Analytics Frameworks and Methods II

Text Analysis of Political Tweets (Democrat and Republican)

Organized by: Gregorio Galletti, Katherine Li, Marsya Chairuna

1 Backgrounds and Data

This project aims to predict the appropriate political parties given published arbitrary tweets and compare sentiment differences between two parties in topics

Background

- Democrat and Republican representatives give the public unfiltered & direct political opinions on Twitter.
- This project aims to
  - Predict political parties given arbitrary tweets
  - Compare general differences in sentiments between the two parties about certain topics

Data

- Tweets from all the representatives (latest 200 as of May, 2018) with following columns:
  - Party name: Democrat (49%), Replican (51%)
  - Twitter Handles
  - Tweets
- 84,502 observations

2 Methodology: Sentiment Analysis

Analyze and identify Republicans and Democrats sentiment about specific topics

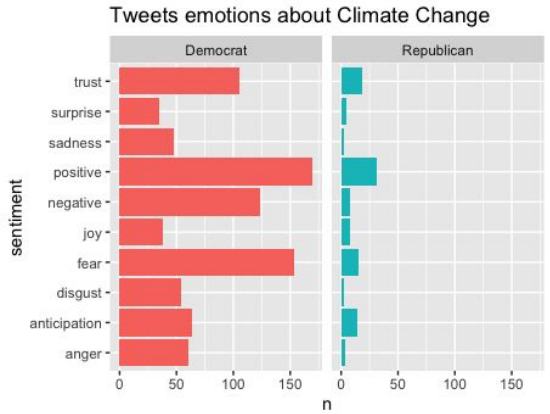
Hypothesis

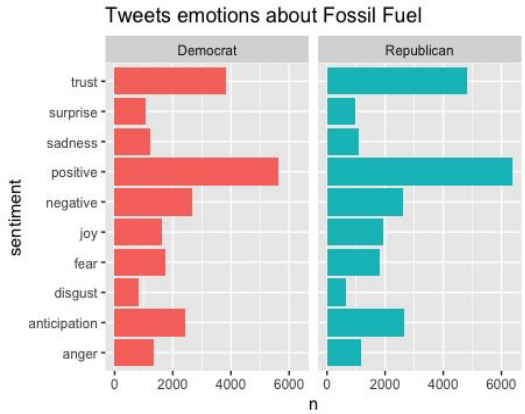
Representatives from the two parties have very different emotions and thoughts about Climate Change and Fossil Fuels

Key Words Filtering

Climate Change: global warming,climate change, ghg, greenhouse gasses
Fossil Fuel: fossil fuels, oil, natural gas, petroleum, ff

Sentiment Analyst Results: H0 rejected





2 Methodology: Classification Model

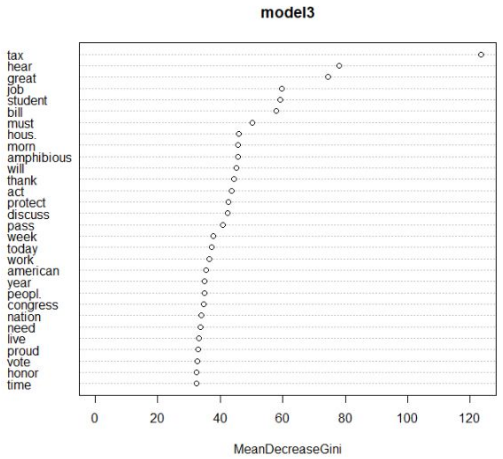
To predict the appropriate political parties using the tweets dataset, we compared the performance of decision tree, Random Forest, and logistic regression

Data Processing

- Preliminary cleaning of text: remove whitespace, remove non-alphabetic characters and special characters, remove URL, punctuations, numbers, stop words
- Stem documents and remove sparse terms, with sparse value of **0.98**
- We retained 57 terms out of 62,259 variables

Models and Parameters Tuning (PCA = 10 Dimensions)

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



3 Summary of Results and Conclusion

Summary of Results

- Sentiment analysis: While 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.
- Classification model: Random Forest is the best-performing model, but the differences between accuracy results for all models are marginal. “Tax”, “job”, and “bill” are among the most important variables

Conclusion

- Recommendation: political parties should better leverage Twitter to promote and their ideology and distinguish their values from other parties.
- Improvement for future research
    - Classification Models
        - More updated tweets that incorporate more topic variety
        - Use reliable political opinions of 2 parties on other social platforms ➔ improve on prediction accuracy
- Sentiment Analysis
    - Conduct analysis in other trendy political topics (such as tax, minimal wages) ➔ gain understanding of sentimental differences of parties from different perspectives