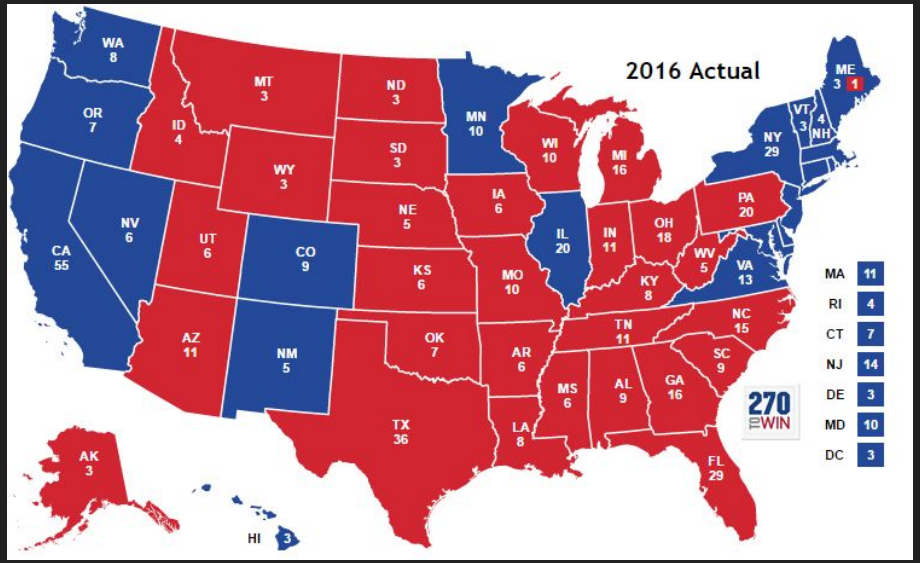
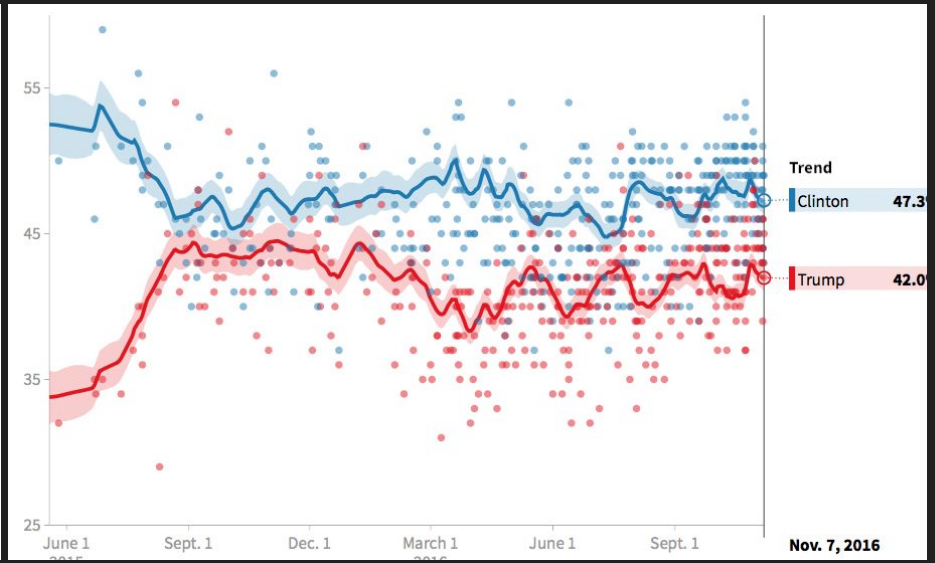
Decide people’s Political Leanings based on their Twitter posts

Team 10
Xinming Liu xil231@pitt.edu
Jingyang Li jil281@pitt.edu
Xuezhi Xu xux15@pitt.edu

Motivation

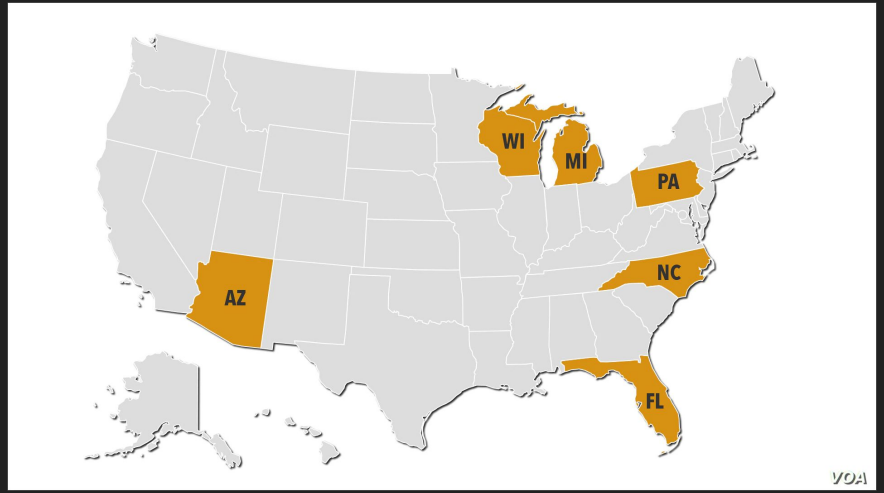
- Are polls very predictive?
- What can we learn from social media posts related to election?





Project Goal

- Predict the political leanings of Twitter users by their tweets during the 2016 US presidential election debates
- Find out which states are potentially more important for the candidate before the election and compare with their road maps back in 2016



Dataset

Set A:

22K users whose party affiliation were identified through their vote registration

Set B:

831K users whose supporting candidates (Clinton or Trump) were identified from their following information

Link to Datasets

Dataset

Set A

users

userID	state_code	party
1	AK	D
2	AK	D

tweets_debate1

userID	text	created_at	favorite_count	followers_account	friends_count	location
11344	Police start the presidential debate memes swirling on Twitter - CNET: It's not only a big night for demo... https://t.co/2DdSv4wuzH	09/26/2016 21:00:00 EDT	1	123	96	East Point, GA

Set B

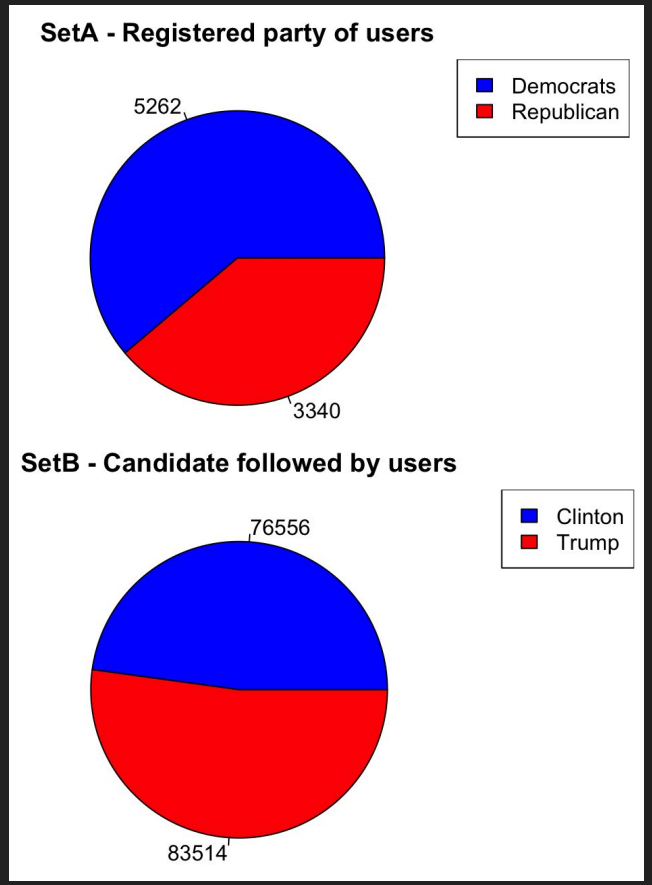
users

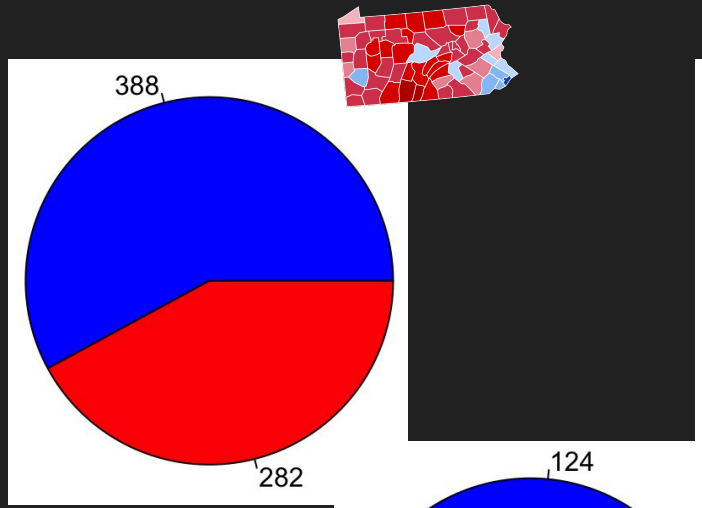
userID	follow_candidate
12846710	NAN
12820601	NAN

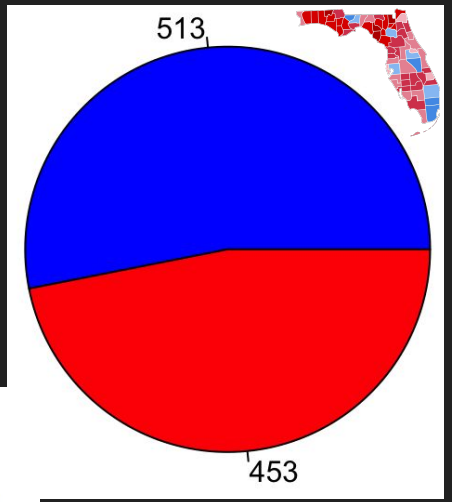
tweets_debateVP

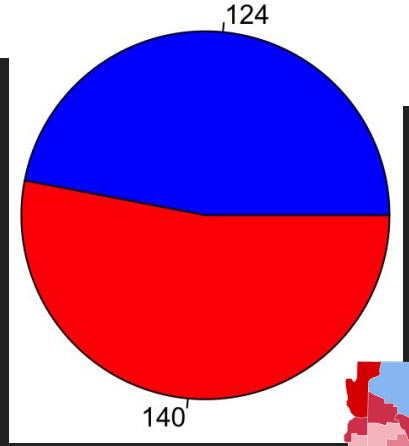
userID	text	created_at	favorite_count	followers_account	friends_count	location
13050965	RT @readbrooks: Sanders: "Look at which candidate is going to stand up for working families, and which candidate is going to stand up for t…	10/04/2016 21:00:00 EDT	66387	446	348	Welcome, NC
682896	RT @LtoldilloM: Former accountant refutes Trump's tax brilliance: I did all the work. #ConDon @realdonaldtrump #VPDebate https://t.co/p.l3K6	10/04/2016 21:00:00 EDT	7894	463	2024	

Data Glimpse









Methods

- Text Mining

  See if tweets of people supporting D/R show some semantic difference

- Clustering

  See if tweets could be clustered into the two parties

- Classification

  Predict political leanings based on tweets

tm v0.7-7

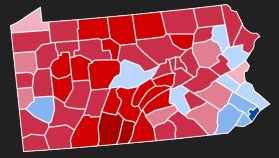
by Ingo Feinerer

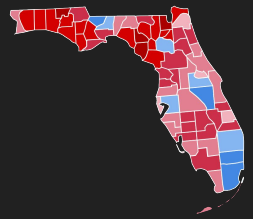
caret v6.0-86

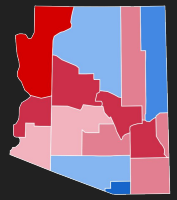
by Max Kuhn

Method 1 – Text Mining

Focusing on three swing states (PA, FL, AZ) across 3 presidential debates + 1 VP debate







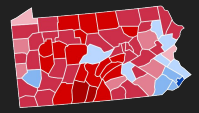
- Preprocess raw tweet text data to build Document-Term Matrix
- Perform visualization based on TF-IDF, LSA, NMF and create a wordcloud (freq>5)

tm v0.7-7

by Ingo Feinerer

Method 2 – Clustering

- Try clustering tweet texts of the whole Set A, Set B, +



- Remove sparse terms in Document-Term Matrix (tune max-sparsity)

- Due to limited computer memory, we only did K-means clustering

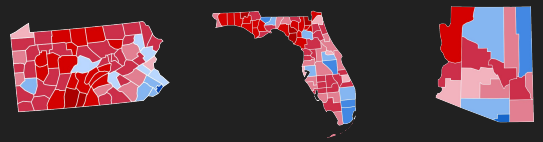
  (k=2 for two parties, k=3 with neutral terms included)

stats v3.6.2

by R-core R-core@R-project.org

Method 3 – Classification

- Try to predict people's political affiliation based on their tweets

  3 presidential debates + 1 VP debate + 

- Use stratified sampling to preserve class balance (Set B)
- Remove sparse terms in Document-Term Matrix (tune max-sparsity)
- Training & Testing different classifiers with 10-fold CV:

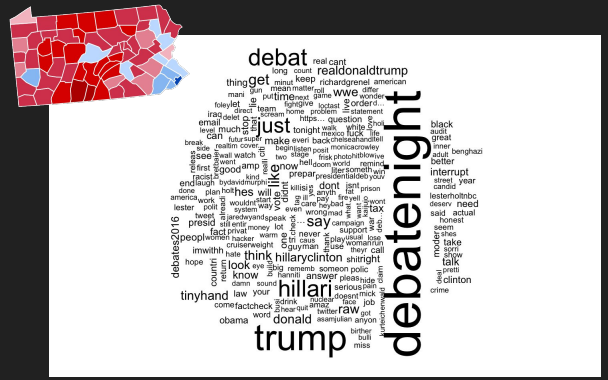
  Logistic Regression + Naive Bayes + SVM (3 types of kernels) + Decision Tree + AdaBoost

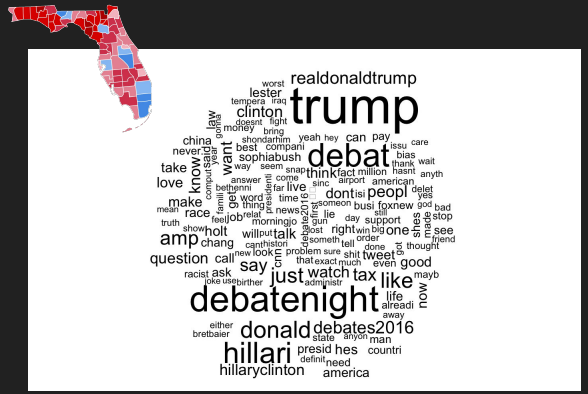
- Evaluation

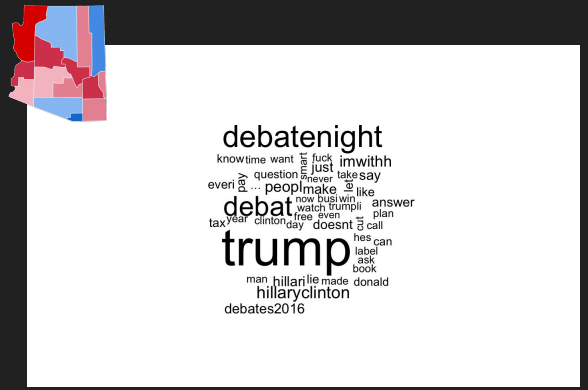
caret v6.0-86

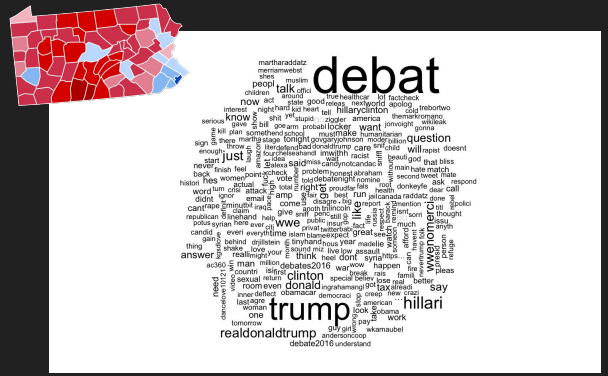
by Max Kuhn

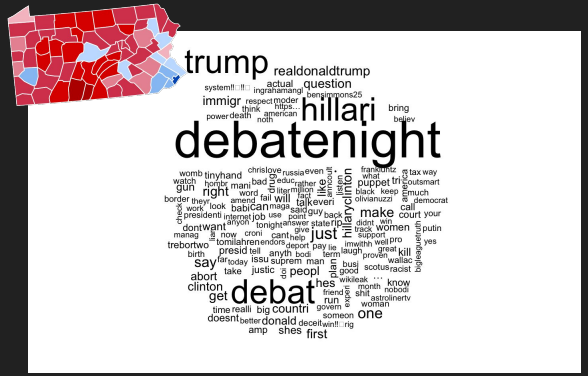
Text Mining Results

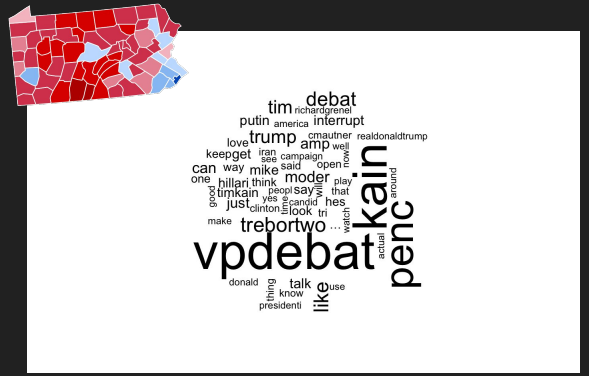




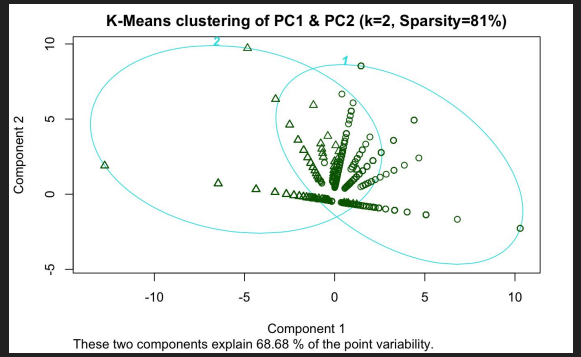


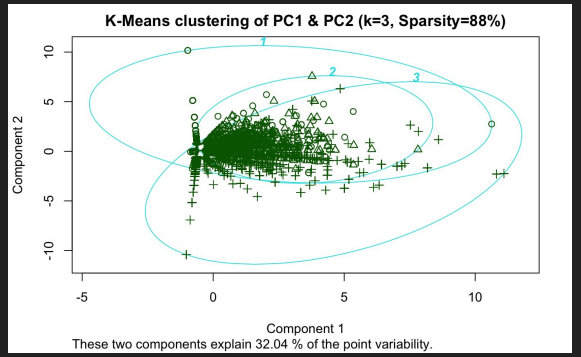


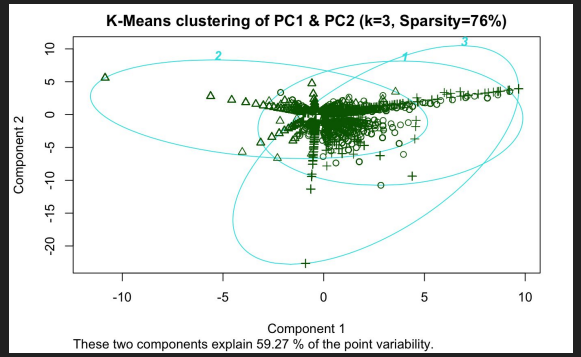


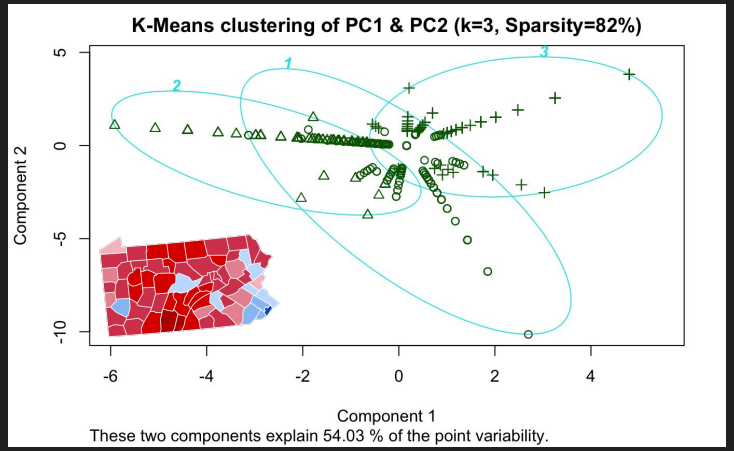


Clustering Results

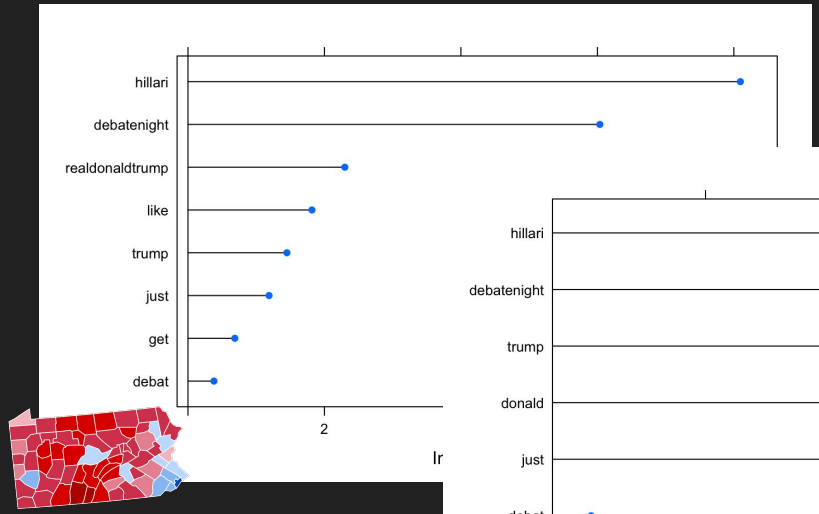


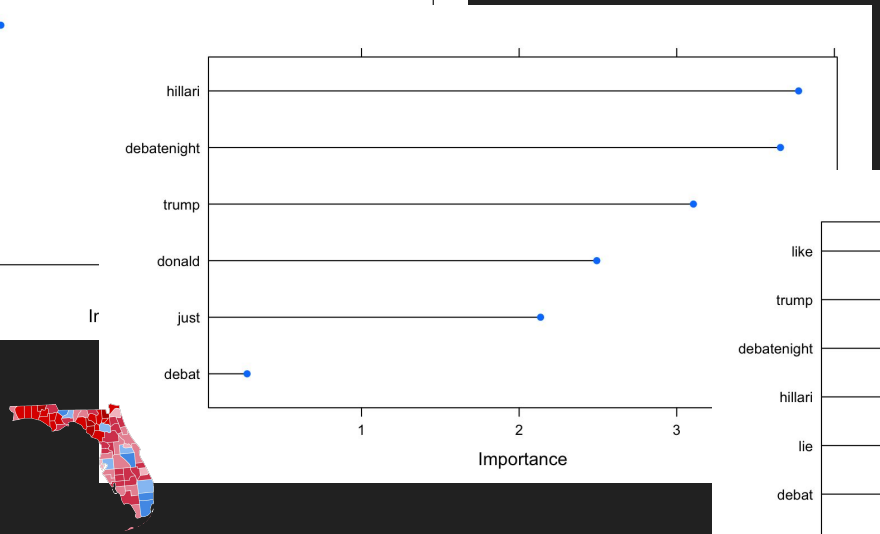


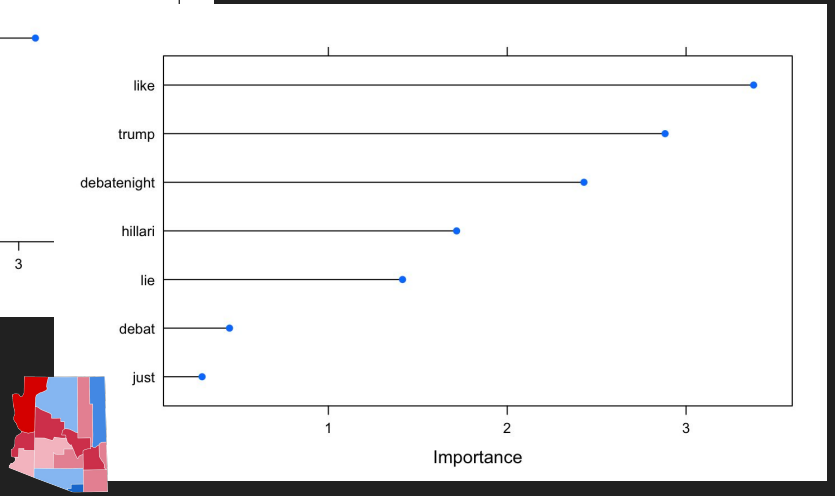




Classification Results – Term Importance

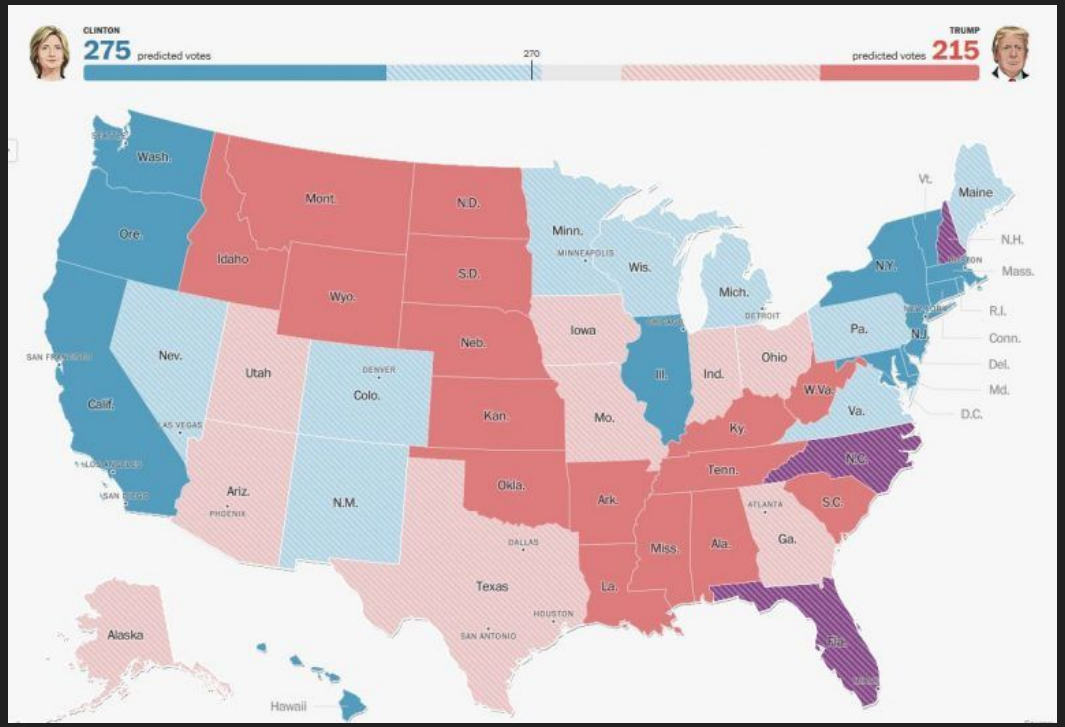






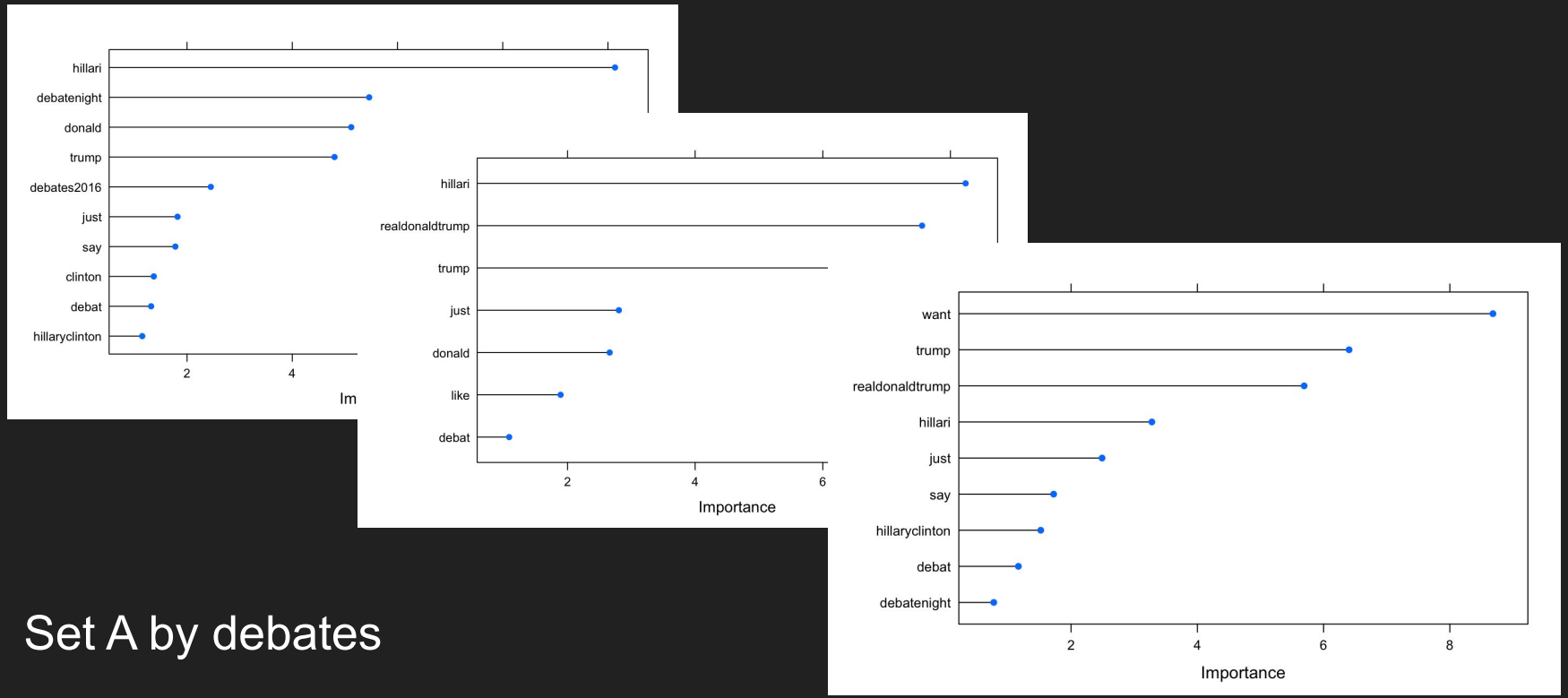
Set A by swing states

Polls after debates and before election



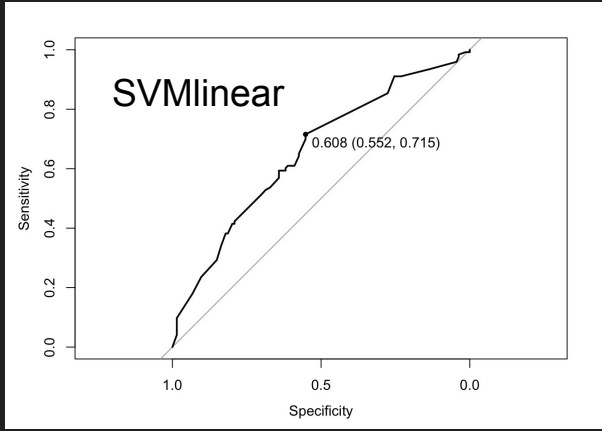
CNN prediction and polls before the election

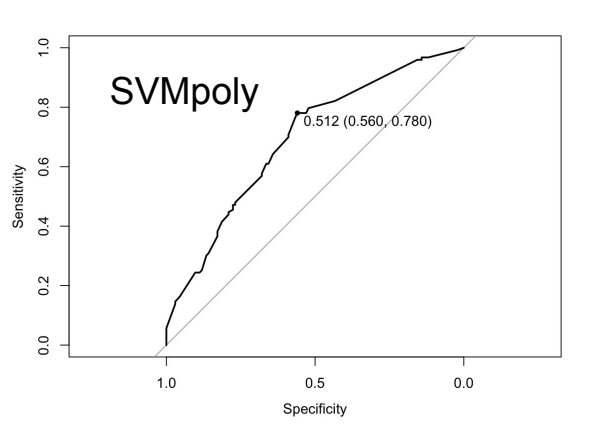
Classification Results – Term Importance

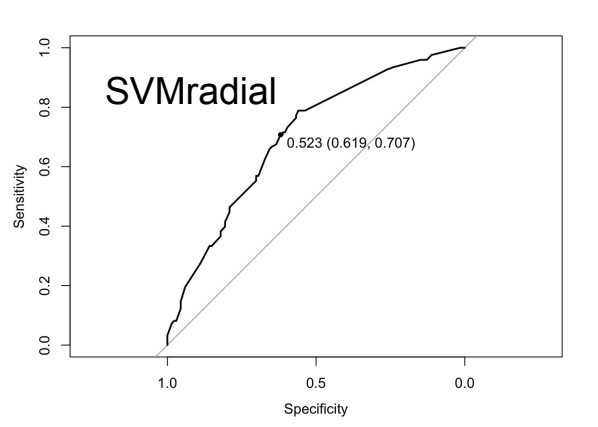


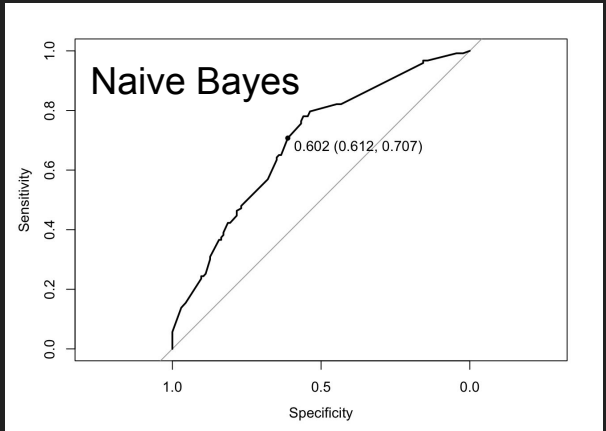
Set A by debates

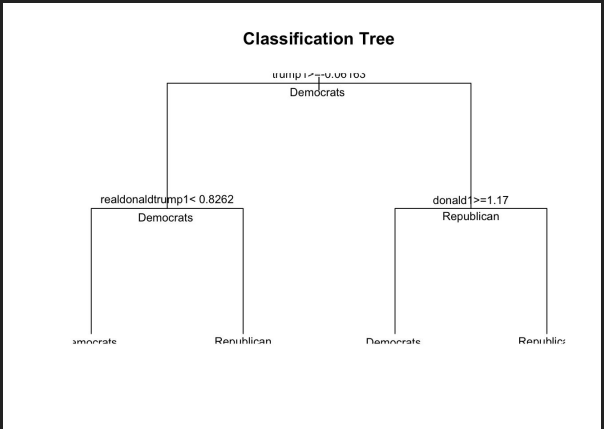
Classification Evaluation – AUC based Set B Debate 3

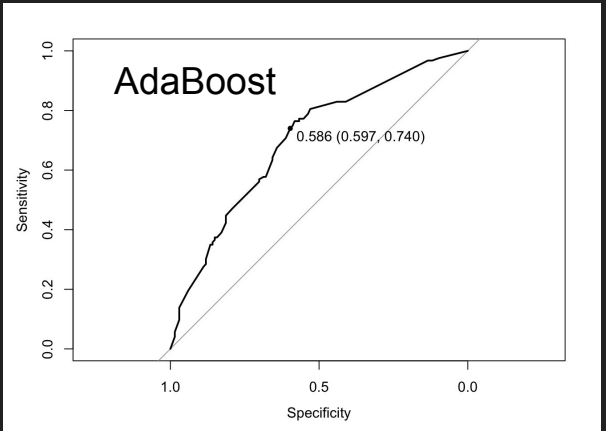




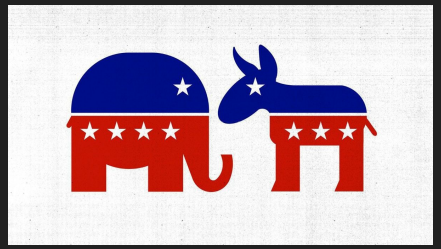








Conclusion – Project Pipelines



- Text preprocessing

  From two raw tables to Document-Term Matrix

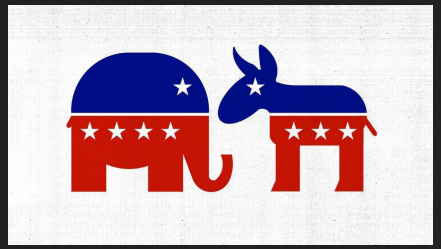
- Data mining

  Text Mining, Clustering and Classification

- Measuring classification performance

  ROC curve / AUC and compare with polls after debates before vote

Conclusion – Problems solved



- Document-Term Matrix too sparse to perform operations

  tm::removeSparseTerms()

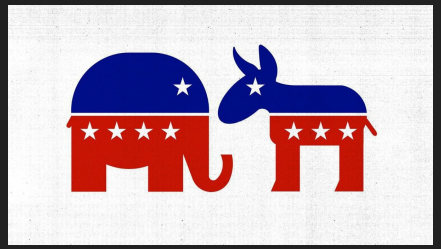
- Set B too large to calculate

  Stratified Sampling

- createDataPartition() not working perfectly

  Manually assign same levels to training and test set

Future Improvements



- Text Mining

  Could potentially work on larger sets if dimensionality reduction applied

- Clustering

  Controlled low-frequency term removal, quantitative evaluation, try Hclust...

- Classification

  Perform LSA, controlled low-frequency term removal, larger tuneGrid / tuneLength...

Thanks for Listening!

Team 10
Xinming Liu xil231@pitt.edu
Jingyang Li jil281@pitt.edu
Xuezhi Xu xux15@pitt.edu