Political Classification via Twitter Feeds

Alex Pham, Darian Martos, Fernando Ramos
Department of Computer Science, Stanford University

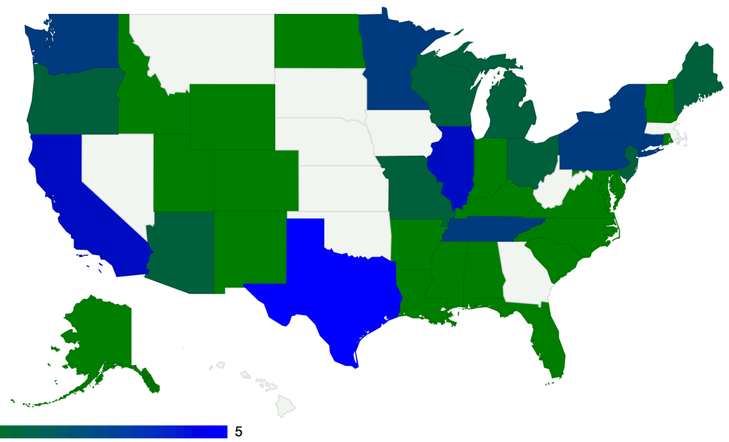
Task & Model

- Task: Predict a person's political alignment given their Twitter feed.
- For our input, we will be using tweets from users who have known scores via GovTrack methodology.
  - GovTrack ideology maps out Members of Congress based on his or her pattern of cosponsorship

Training Methods

- Bag of Words — Unigram
  - Multinomial Naive Bayes
  - Logistic Regression
  - Support Vector Machines
- Our feature vectors were sparse vectors that represented word count

Number of Politicians Used for Training



Evaluation

- We used cross validation to see how closely we could predict politician's political alignment.
- We are still in the processing of improving our metrics because of updates to our project.
- We are currently only testing and training on tweets from politicians because we do not have scores for other users.

Results

Naive Bayes

- Trained on 200,000+ Tweets, Tested on 40,000+ Tweets
- Results below are for different folds

Folds with No Buffer		Folds with Buffer of One	
1	0.619857	1	0.734951
2	0.616320	2	0.735413
3	0.613830	3	0.736589
4	0.613119	4	0.723618
5	0.608167	5	0.721840
6	0.610334	6	0.737398
7	0.617619	7	0.735570
8	0.617619	8	0.724857
9	0.612828	9	0.735434
10	0.613203	10	0.729249

Support Vector Machine & Logistic Regression

- Trained on over 200,000 Tweets
- When tested on 15 separate profiles, SVM correctly predicted 12/15 profiles with a buffer of two and LR correctly predicted 11/15 profiles with a buffer of two

Interesting Data

Top Three Words by Class

1	dc, bill, trump	6	happy, birthday, tax
2	cbrangel, great, day	7	il, great, bill
3	mepolitics, great, health	8	great, house, bill
4	great, bill, work	9	house, jobs, great
5	great, bill, county	10	great, house, obamacare

Words & Number of Occurrences

clinton	576	trump	5429
healthcare	1684	obamacare	3284
bill	9998	america	10720

Discussion & Conclusion

- Even with bag of words/unigram models, accuracy is relatively high
- Differences exist between binary and multi-labeled classification
- Further work to be done:
  - Implementing Random Forest
  - Comparing across bigram/trigram models
  - Comparing with smaller and larger training and test data sets
  - Natural Language Processing methods: word filtering, segmentation, Laplace & other smoothing