CS 583 Project 2: Twitter Data Mining

Yaru Shi & Lei Zheng

University of Illinois at Chicago

December 4, 2015

Data Pre-processing

- Data cleaning
  1. Removed "mixed" category data;
  2. Filtered missing-label data;
  3. Discarded labels out of the range, e.g. !!!.
- Features extraction
  1. Length of tweets are limited: 140 characters;
  2. Not enough information if we put too many words in one bag;
  3. Unigram is chosen.
- Vectorization:
  1. Transformed the data into numerical feature vectors;
  2. Re-weighted data: tf-idf

Classification and Results

- blue represents Obama; red represents Romney

Algorithms	Results						
	Positive class			Negative class			Overall
	p	r	F	p	r	F	
Naive Bayes	0.64	0.45	0.52	0.50	0.49	0.49	0.48
	0.46	0.27	0.34	0.43	0.61	0.49	0.42
Logistic	0.53	0.64	0.58	0.62	0.57	0.58	0.57
	0.24	0.61	0.34	0.86	0.58	0.68	0.55
SVM	0.59	0.61	0.59	0.61	0.58	0.57	0.57
	0.38	0.54	0.44	0.77	0.63	0.68	0.57

Next Steps

- Data Balance: Tweets voting for Romney is half less than those are against him.
  Solution: Augment data
- Feature Selection: LASSO, Forward/Backward feature selection, etc;
- Parameter Tuning: kernel function, penalized parameter, penalty methods.