2016 Presidential Election Prediction using Twitter

Weifeng Li
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Abstract

Nowadays, data of social medial websites are getting more and more popular to be used as one of the most important data source for the data mining from which we can find the useful and interesting patterns. In this project, base on twitter data set that I collect using twitter API, I performed the sentimental mining and topic modeling. In the data collection phase, I used keywords such as the candidates' name to filter the related data decreasing the noise to the most extend. To accomplish the sentimental mining, I chose Naïve Bayes algorithm and Support vector machine Model(SVM) two of the most commonly used algorithms that can be used as the classifier in the sentimental analysis. Then I trained these classifiers using a data set which was also from twitter and was related to 2016 presidential election from Kaggle and made the predication using twitter data set that I collected. Besides, Latent Dirichlet Allocation model was used to fulfil the topic modeling analysis finding the most frequent topics from the data of presidential election related tweets. At last, I evaluated the performance of each classification algorithm.

Problem Definition

Firstly, using sentimental analysis, based on the tweeter data, the popularities of each candidate will be come up with. In this phase, two calcific algorithms Naïve Bayes and Random Forest will be used as the classifier and performance of them will be evaluated at the end.

Secondly, I will compare the results of the analysis of tweeter data with the results of polls and contributions to see if they are comparable and if they roughly have the same trend.

Besides, I will use Latent Dirichlet allocation(LDA) model extracting frequent topics related to 2016 presidential election using tweeter data set. For better illustration of the results, I will visualize the results in World Cloud.

Design and Implementation

To find the best parameters for these classification algorithms, I used cross-validation method for the parameter tuning. As can be seen below, throw this way, the best kernel parameter for SVM and the best alpha parameter for Naïve Bayes were found.

PARAMETER TUNNING RESULT(SVM)

kernel	C	gamma	Score
rbf	1	0.001	0.5
rbf	1	0.0001	0.5
rbf	10	0.001	0.5
rbf	10	0.0001	0.5
rbf	100	0.001	0.704
rbf	100	0.0001	0.5
rbf	1000	0.001	0.803
rbf	1000	0.0001	0.704
linear	1	auto	0.812
linear	10	auto	0.795
linear	100	auto	0.793
linear	1000	auto	0.793

PARAMETER TUNNING RESULT(NAIVE BAYES BAYES)

alpha	Score
0.10000000000000001	0.781
9.9999999999999995e-07	0.773
3.5938136638046257e-06	0.774
1.2915496650148827e-05	0.774
4.6415888336127818e-05	0.774
0.00016681005372000591	0.773
0.00059948425031894088	0.770
0.0021544346900318843	0.768
0.0077426368268112772	0.766
0.02782559402207126	0.771

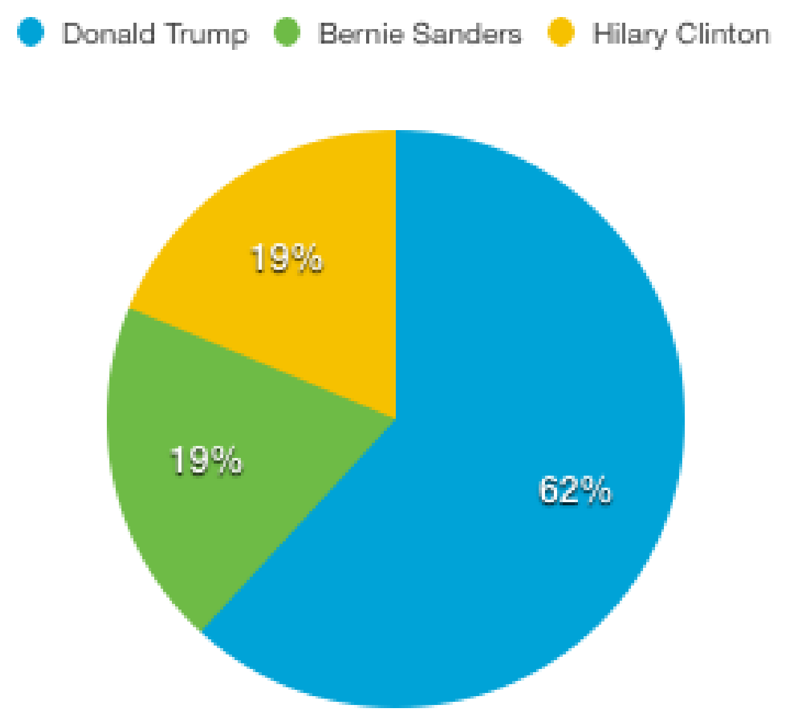
Result

We can see that several interesting things from the result. Firstly, the overwhelming number of tweets are Negatives. Secondly, Donald Trump was more popular than the other candidates. Last, although Trump was more popular, Hilary was more possible win the election.

Tweets Distribution

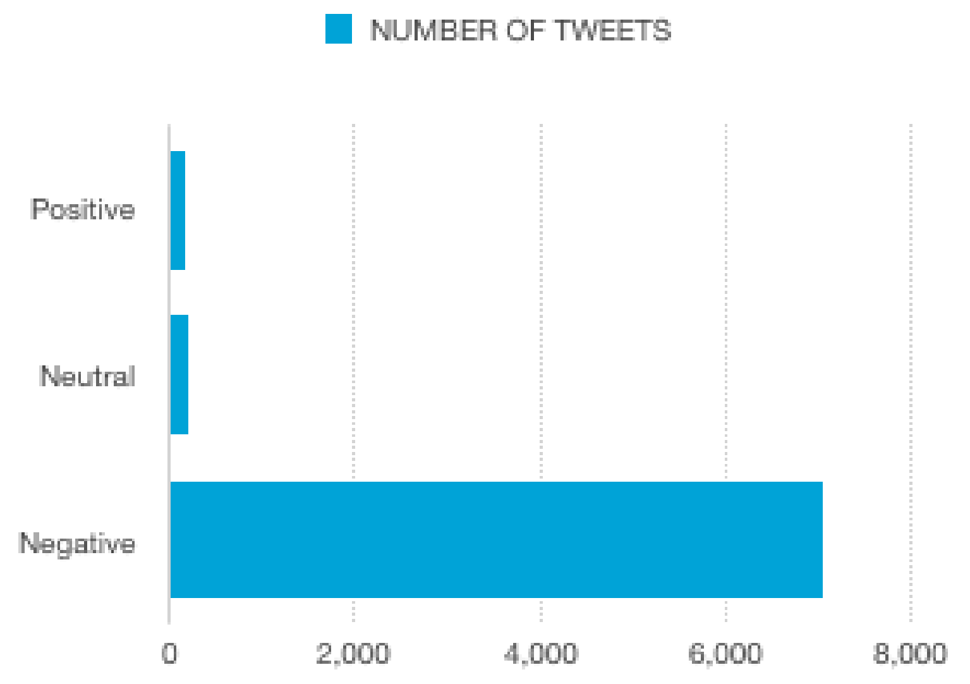
CANDIDATE	NUMBER OF TWEETS
Donald Trump	4,576
Bernie Sanders	1,443
Hilary Clinton	1,386
TOTAL NUMBER	7,405

NUMBER OF TWEETS BY CANDIDATE



SENTIMENT	NUMBER OF TWEETS
Positive	160
Neutral	198
Negative	7,047
TOTAL NUMBER	7,405

NUMBER OF TWEETS BY SENTIMENT

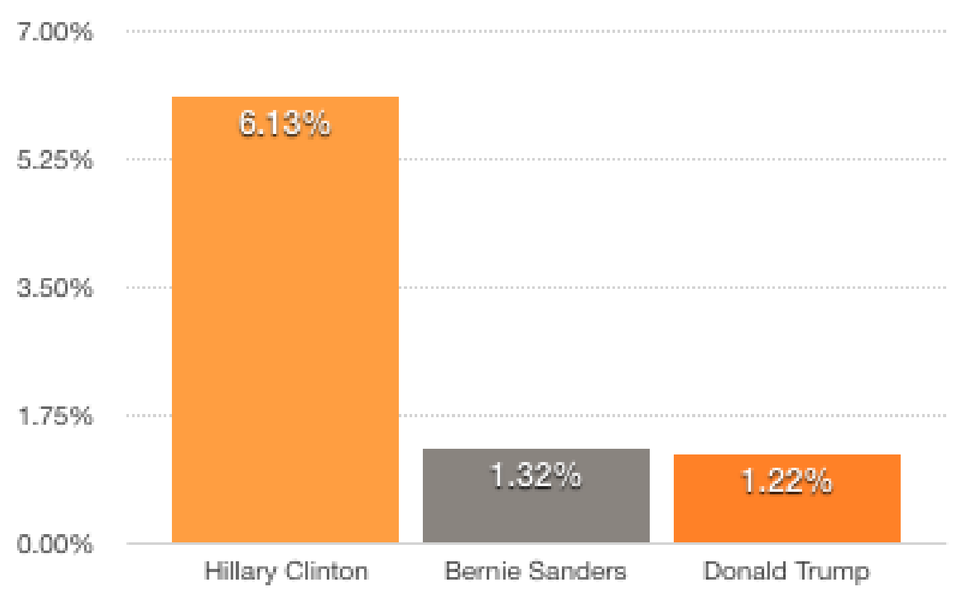


Supporting Rate By Candidate

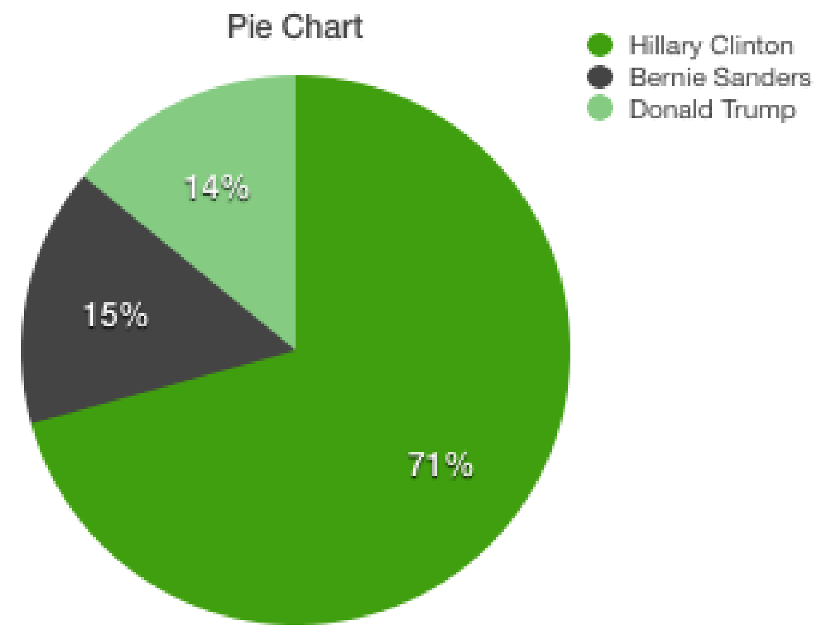
Positive Rate by Candidate

CANDIDATE	RATE OF POSITIVE
Hillary Clinton	6.13%
Bernie Sanders	1.32%
Donald Trump	1.22%

Column Chart



Pie Chart



Performance Evaluation

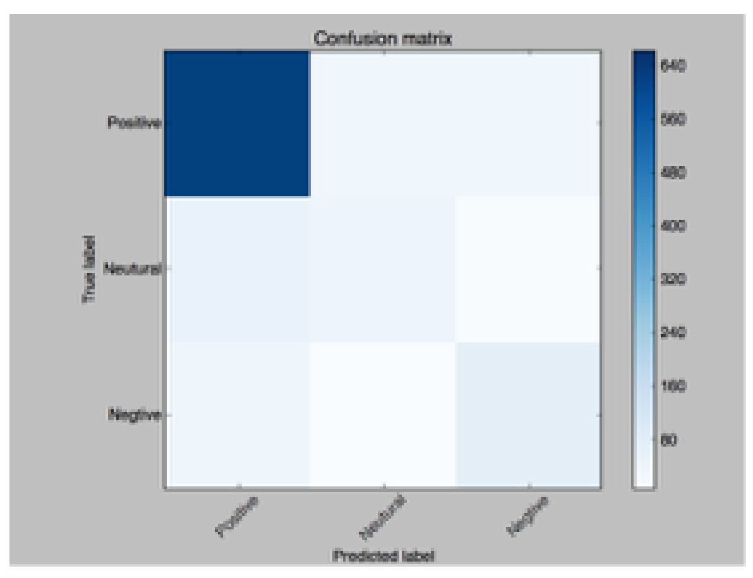
To evaluate the accurate performance, Confusion Matrix was used here. As shown below, the accuracy of Linear SVC classifier was better than Naïve Bayes.

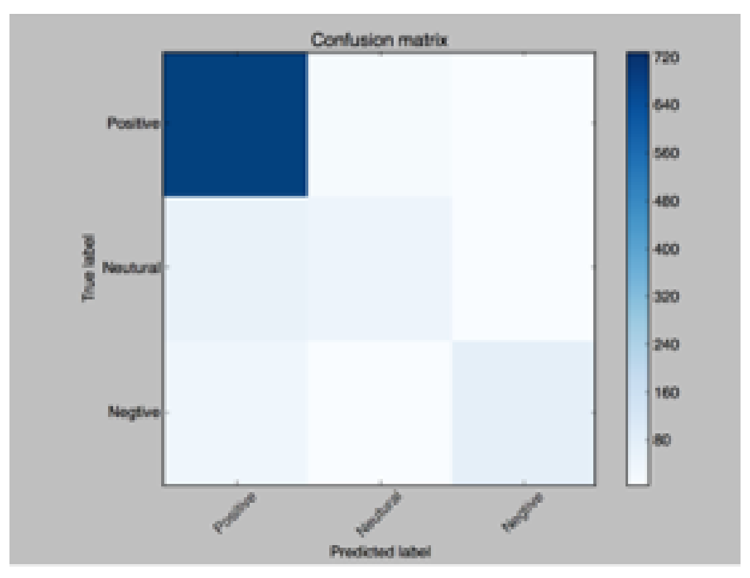
CONFUSION MATRIX(NAIVE BAYES)

	PRECISION	RECAL	F1-SCORE	SUPPORT
POSITIVE	0.85	0.87	0.86	758
NEUTURAL	0.52	0.43	0.47	138
NEGTIVE	0.61	0.62	0.61	149
AVG/TOTAL	0.77	0.78	0.77	1045

CONFUSION MATRIX(LINEAR SVC)

	PRECISION	RECAL	F1-SCORE	SUPPORT
POSITIVE	0.86	0.97	0.91	758
NEUTURAL	0.68	0.40	0.50	138
NEGTIVE	0.86	0.63	0.73	149
AVG/TOTAL	0.83	0.84	0.83	1045





Conclusion

As social media such as twitter has not been systematically studied yet, with the development of science and technology especially in big data and machine learning area, more and more new methodologies and algorithms can be used to find the interesting patterns from social media. I use semantic analysis and topic modeling methods mining the twitter data set and find the similar result comparing with the result from other data set such as polls and final contribution data set. The result show that, we can find more useful and interesting pattern and information from social media data mining.