prediction about presidential campaign becomes possible in the generation Z

Insert Subtitle Here

Xinming Liu
xil231@pitt.edu

Xuezhi Xu
xux15@pitt.edu

Jingyang Li
jil281@pitt.edu

ABSTRACT

So far, Twitter has 141 million American daily active users, more and more people tend to post their political views on twitter to express their political opinions or demands. As tweets increase, there are more texts for people to analyze the political leanings of the people who post tweets. Actually there are many sucessful cases to predict people's action based on their user data, like in amazon, there is a catagory that can guess what you like and recommand this for you, we think this use the technology of data mining which provideds inspiration for our project. In this project, we are going to do clustering and classfication to predict people's political leanings. We will use term frequency-inverse document frequency matrices (TF-IDF), Latent semantic analysis (LSA), non-negative matrix factorization (NMF) to build and evaluate our model.

KEYWORDS

Clustering, Classification, Text preprocessing

1 INTRODUCTION

We are now in the era of Z generation which means that people are surrounded by many social apps like Facebook and twitter. The social media has gain attention in the political area. Researchers started at analyzing data on twitter and use tweet content as a basis to analyze people's political leanings. This analysis of the people's political tendencies is very helpful to the election strategy and focus. "Social media services have become areas of political communication. Politicians integrate them in their campaigns, journalists use them as sources and topics, and the public uses them for the discussion of politics." [1] Many people have an opinion that people tend to speak their heart on the internet compared to traditional media express. Many people may be too shy to express their real opinions. Analyzing the tweet content on twitter can easily solve this problem. Analyzing data on social media provides political campaign team a new way to adjust their campaign strategies. Analyzing the data on social media is very important.

We will use the tweets contents during the three debates between two Presidential candidate and the vice-president debate to predict the leanings of people who post tweets. We will build a model and use classification to predict people's leanings and will use clustering methods to get people' political affiliation labels. Since the main two political party are Democratic Party and Republican Party, and the two Presidential candidates belong to these two parties, so we will only use the tweets content of people who support Democratic Party and Republican Party. As we all know, there are some states that will support one specific party most of the time, like Most people in New York and California will support Democratic Party mostly and Most people in Texas will support Republican party mostly, so we will exclude these states and reserve some state called "swing state". These "swing state" may vote for Republican party or Democratic party, they both have chance, so we choose the Arizona, Florida and Pennsylvania as our main research object.

2 RELATED WORK

There are many different kinds of research on analyzing the tweets about the presidential election. Most of them are about the text mining. Researchers use the content of tweets people post on the internet to conduct their political affiliation. Tweeter has more interactions with political. Like there is a real time commentary, and researchers can use the real-time commentary to study people's political affiliation. "Real-time commentary, also called 'live-tweeting', has become a widely used feature of the social networking site Twitter." [2] "One such application is in the field of politics, where political entities need to understand public opinion and thus determine their campaigning strategy. Sentiment analysis on social media data has been seen by many as an effective tool to monitor user preferences and inclination. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis" [3].

In this project, we will do classification to decide people's political affiliations. We will use the text mining to classify people. The topic people post on the twitter may represent their political affiliation. For example, leftists in the Democratic camp may focus more on the environment or human rights movements. However, for the republican rightists, they may more likely to pay attention on taxation and employment opportunities. We use this difference to predict people's political affiliation. We also use the emotions that people have in their tweet topics (they support it or oppose it). We will also use text mining to do clustering. We try to make people with similar political opinions in the same group.

3 DATASET

The data we use is provided by the Professor Yuru Lin in topic3(the tweets from two sets of Twitter users during the four candidates debates in 2016 United States presidential election). We think that the analysis of swing states has more value than other state in the research of prediction of the results of presidential elections.

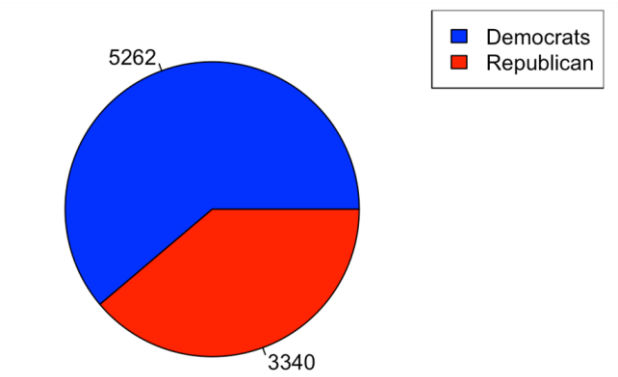

Figure 1. Registered party of user setA

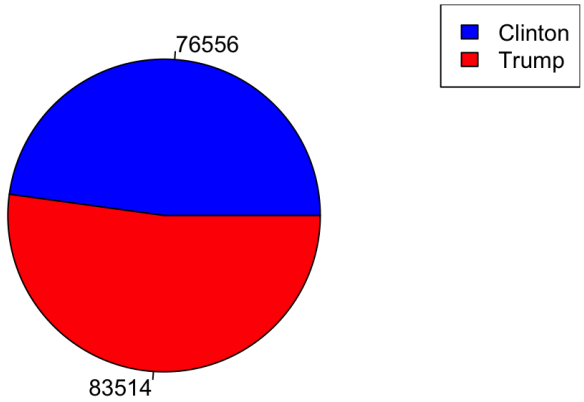

Figure 2. Registered party of user set B

Since the main two political parties are democratic party and republican party, so we only preserve the user data which belongs to democratic party and republican party. From the preprocessing the data, we clean the data, remove the whitespace and extra the word term. Then we build the document-term matrix. we found a problem is that dataset B is too large for us to do the classification. The dataset is too large that we can't train the model with the dataset B. We choose serval swing states as our dataset. The swing states are PA, FL and AZ.

4 METHOD

4.1 Text Mining

In the text mining part, we use TF-IDF (term frequency–inverse document frequency), LSA (Latent Semantic Indexing) and NMF (Non-negative matrix factorizatio).

First, we apply TF-IDF (term frequency–inverse document frequency) on the corpus. Figure 3 show the performance of the model

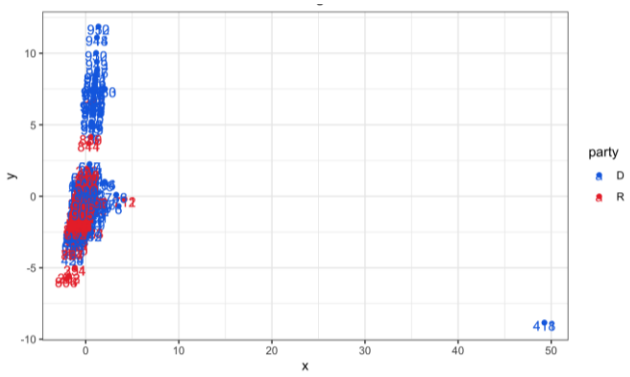

Figure 3. Debate 2 PA clustering based on TF-IDF

Then we apply LSA (Latent Semantic Indexing) on the corpus. Figure 4 show the performance of the model

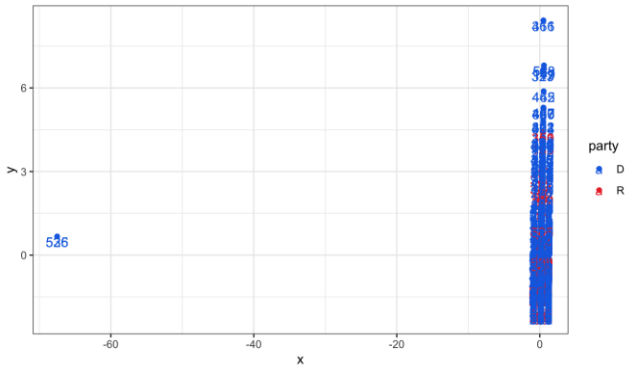

Figure 4. Debate 2 FL clustering based on LSA

Then we apply NMF (Non-negative matrix factorizatio).
on the corpus. Figure 5 show the performance of the model

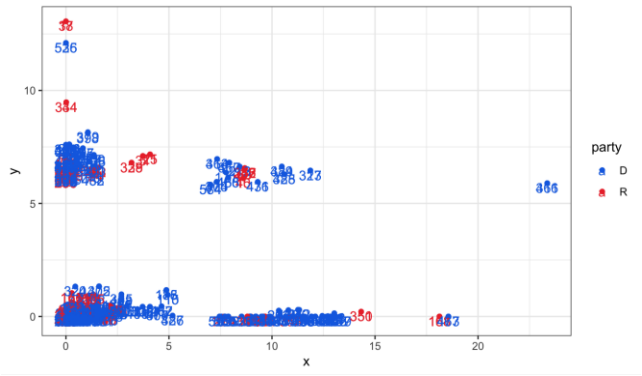

Figure 5. Debate 2 FL clustering based on NMF

Finally, we get our text mining results

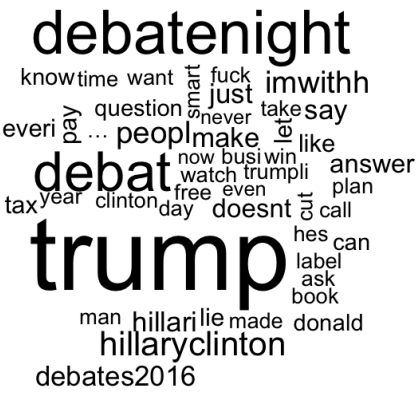

Figure 6. Debate 1 PA Text mining results

We did every debate from debate 1 to debate VP in set A in 3 swing states and only swing state for set B. The complete results are in our github project. As shown in the Figure 6, the word "trump" appears most times which means this word is a popular word after the debate 1 on twitter. We choose the word which appearance frequency is above 5 times. This is for us to analyze the topics in the tweets, to help us to find words that we need to study accurately.

4.2 Classification

In this part we extract the tweets' text and user's party whose party belongs to Democratic Party or Republican party. Then we clean the corpus which base on the data we extract. Then we build the document matrix and remove the low-frequency word which frequency is below 0.05. When we set the sparsity to 0.97, which means that we filter the word frequency which below 0.03, there are 17 terms left, which is too much. Finally, we set the sparsity to 0.95 and only 5 terms left which is fine to us. Then we build the data frame which contains DocumentTermMatrix and tweeter user' political parties. Then we choose the 10-fold cross validation with 70% data for training and 30% data for testing. We did this in the swing states and the whole country in data set A and data set B. The first method we use is KNN (K-nearest neighbors), then this method brings many ties which can't show the results clearly, so we abandon this method. We also try to use pSVM method to train the model, however, due to the CPU of the computer limitation, we can't apply a grid to tune parameters.

Finally, we use lSVM, rSVM, NB(naïve bayes), LR(linear regression), DT(decision tree), ADA to train the model. We use the test data DocumentTermMatrix and the prediction DocumentTermMatrix to plot the ROC curve. Here are the results.

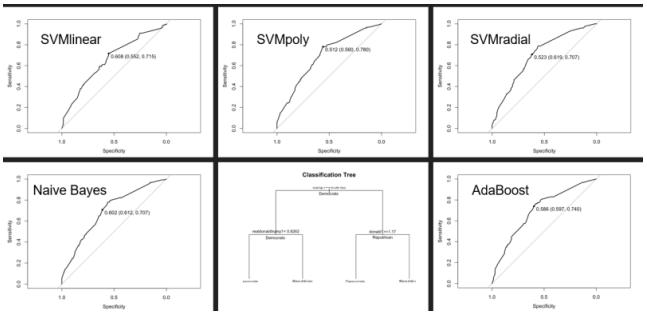

Figure 7. ROC curve on different methods

We did the classification in every swing state and the whole country for both data set A and only swing state for data set B, we store the complete results in our github project.

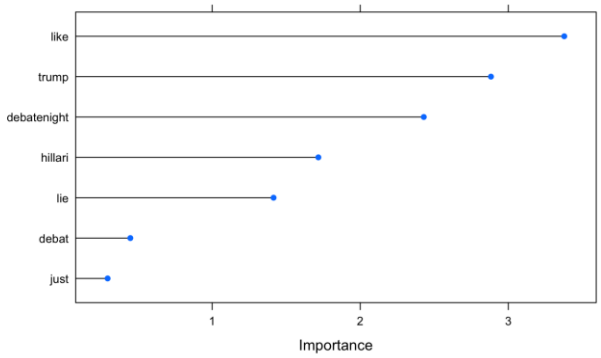

Figure 8. Word importance in Arizona

4.3 Clustering

In this part, we create our corpus and DocumentTermMatrix due to our limited memory, we apply TF-IDF weighting and do aggressive feature words extraction, then we cluster the data, due to our memory limitation, we can only do k-means with centers equals 2 and nstart equals 10. We found that the maximum sparsity equals 90%, there will only one word left, it's the #Debate night. We do the visualization of two most frequent terms, K-Means clustering of PC1 & PC2 and . K-Means clustering of PC1 & PC2(Sparsity=88%). All words in the visualization are in the stem type, for example, "debat" is actually the stem type of the word "debate". We do clustering in swing states (PA, FL, AZ) and the whole country for both data set A and only swing states(PA, FL, AZ) for data set B. The complete results are in our github project.

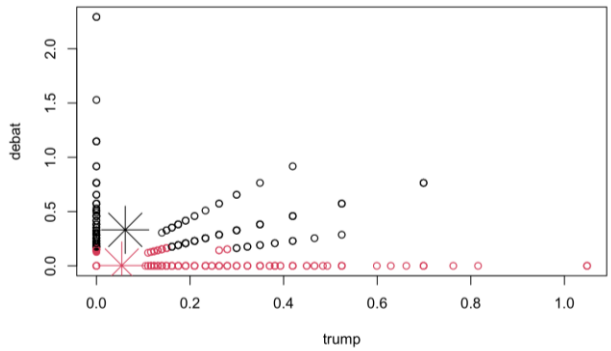

Figure 9. visualization of two most frequent terms(Sparsity=78%)

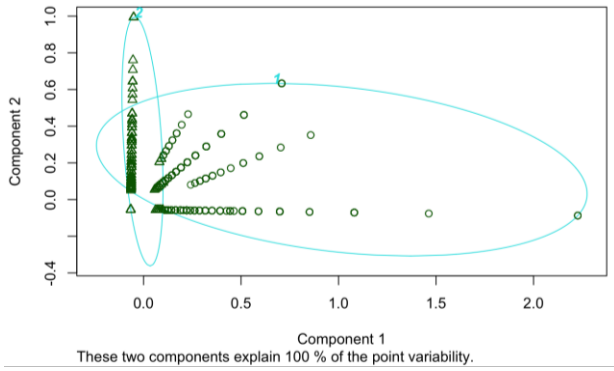

Figure 10. K-Means clustering of PC1 & PC2(Sparsity=78%)

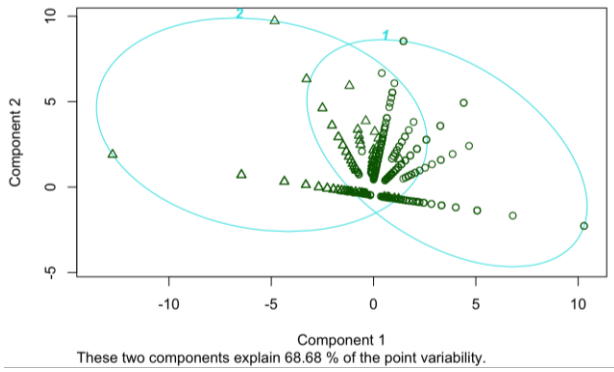

Figure 10. K-Means clustering of PC1 & PC2(Sparsity=78%)

5 EVALUATION

In the state of Arizona, we compare different performance with different models.

Model performance	lSVM	rSVM	NB	LR	DT	ADA
accuracy	0.502	0.485	0.585	0.585	0.585	0.597
threshold	0.712	0.687	0.510	0.747	0.675	0.796

Table 1 model performance

Here are the results of the different methods in ROC curves

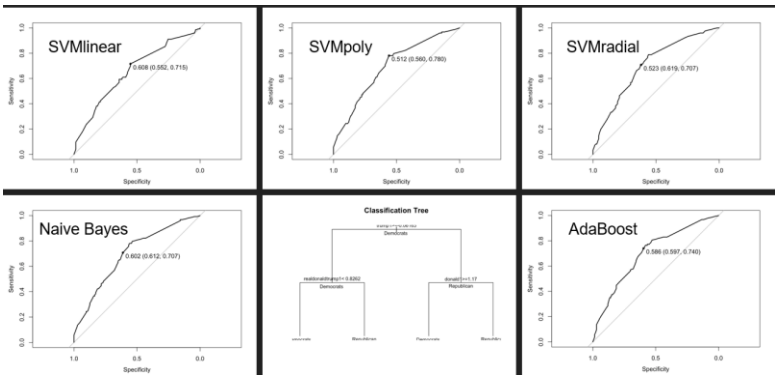

Figure 7. ROC curve on different methods

Based on the ROC curve and the performance table, the AdaBoost model has the best performance compared with other models.

6 DISCUSSION

We also met many problems. First of all, the dataset is too large. We can't train the model with data set B in the whole country. We want to upgrade our computer hardcore and then run the model or potentially we want to find a method to do dimensionality reduction. Then maybe we can perform the large data set. Next improvements we want to do in the future is about the clustering. We have difficulty in sparse word removing in the part of clustering since once we apply the sparse removal on it, there will come out many NA values which will influence our clustering, so we want to control the low-frequency removal and try quantitative evaluation and Hclust so we can have more ways to compare which one is good. For the classification we want to perform LSA and try to solve large tune grid problem. One other thing we find is that in tweeter, it is actually hard to find the user's political party which means that we can't know user's political party in advance, so it's maybe hard to train the model and in reality, people seems tend to post tweets that is unrelated to the political topics, people may tend to criticize some politicians but few of them will post real political issues like tax or

energy policy which will make the prediction more difficult. These are the problems we want to solve in the future.

7 CONCLUSION

In text preprocessing, we create Document-TermMatrix from two raw tables. Then we applied text mining, clustering and classification. Finally, we measure the performance of different models in classification, then choose the ADA model as the best model.

ACKNOWLEDGMENTS

We want to thank Professor yuru Lin for giving us the knowledge of data mining, such as word processing, etc… which we can use in this project, and for her patient guidance during the class.

REFERENCES

[1] Jungherr, A. (2014). The Logic of Political Coverage on Twitter: Temporal Dynamics and Content. Journal of Communication, 64(2), 239–259. https://doi.org/10.1111/jcom.12087

[2] Kjeldsen, L. (2016). Event-as-participation: building a framework for the practice of "live-tweeting" during televised public events. Culture & Society, 38(7), 1064–1079. https://doi.org/10.1177/0163443716664482

[3] Ramteke, Godhia. "Election Result Prediction Using Twitter Sentiment Analysis." 2016 International Conference on Inventive Computation Technologies (ICICT). Vol. 1. IEEE, 2016. 1–5. Web.