Tweet Analysis: Presidential Election 2016

NISARG SHAH

Abstract

 Elections are becoming an increasingly trending topic nowadays with their forthcoming in November 2016.

 Social media is becoming an effective medium for the candidates to convey their message to the voters. Also public can express their opinion of their respective candidate through social media. So it has become a hot topic among researchers to understand the public opinion and predict the President of the upcoming election based on public opinion.

 For this project we analysed Twitter data.

We collected around 2000 tweets for each city spanning across 15 cities all around US. Our study shows that Hillary Clinton is the more popular in most cities whereas Ted Cruz is least popular.

Introduction

 Elections give the citizens of any country a chance to elect their choice of leader. For a democratic country elections are very important as it gives a right to every individual of the country to choose their leader. It is the voice of the people which gets listened.

 The social media is becoming an increasingly important medium for the people to raise their voice. Through the social media people express their feelings about each candidate which makes other people aware of how their leader should be.

 Due to the upcoming elections in November 2016 it has become a trending topic among the social media to discuss about the political campaign and this has

motivated us to choose this topic. As the election approaches the future leader will be influenced by public sentiment and opinion. We are analysing the tweets of the top 4 candidates: Donald Trump, Hillary Clinton, Bernie Sanders and Ted Cruz.

 Our main goal for this project is to solve the problem of understanding the current feeling towards each candidate based on public opinions and predict the outcome for the most popular and the least popular candidate in each region.

 Entire project was done as follows: First we collected the tweets from 15 cities spread across entire USA for each of the four candidates. Then we calculated the positive and tweets for each candidate with the help of sentiment analysis. Then the entire data was separated into training and testing data and predictions were made for the testing data on the basis of training data by using various algorithms such as SVM, Logistic Regression, and Naïve Bayes.

Scope

 The main scope of this project is to predict the winner of the upcoming elections based on people’s opinion and sentiments.

 This can be achieved by using various algorithms and getting the accuracy from each algorithm. The algorithm with the maximum accuracy is used for predicting the outcome of all cities.

Related Work

**Data Collection**

 At first data was collected. We selected the cities from each coast as well as

central region so that we could get an overall idea from each region such as which region supports which candidate and which candidate is popular in which region. This data was collected periodically so that it becomes possible to get perfect prediction results.

 We used Rest API to collect data. We selected the following cities from the following states for collection of tweets:

 States (Most populous city):

1. New York - NYC

2. Massachusetts - Boston

3. Illinois - Chicago

4. California - LA

5. Texas - Houston

6. Georgia - Atlanta

7. Florida - Jacksonville

8. Washington - Seattle

9. Washington DC

10. Wisconsin - Milwaukee

11. South Dakota - Sioux Falls

12. Colorado - Denver

13. Arkansas - Little Rock

14. Kentucky – Louisville

15. Idaho – Boise

 Also we used the following query for data collection:

Donald OR Trump OR Hillary OR Clinton OR Bernie OR Sanders OR Ted OR Cruz OR President OR Election 2016 OR Republican OR Democrat OR Vote OR Campaign OR makeamericagreatagain OR feelthebern OR imwithher OR cruzcrew

 This keywords helped to collect the data across each city. For example to collect the tweets in New York City for each candidate we used the following queries:

 Trump: Donald OR Trump OR President OR Election 2016 OR Republican OR Vote OR Campaign OR makeamericagreatagain OR politics OR debate OR poll OR delegate OR voters

OR campaign OR caucus OR candidate

OR Presidential OR DonaldTrump

 Hillary: Hillary OR Clinton OR President OR Election 2016 OR Democrat OR Vote OR Campaign OR imwithher OR politics OR debate OR poll OR delegate OR

voters OR campaign OR caucus OR candidate OR Presidential OR HillaryClinton

 Ted Cruz: Ted OR Cruz OR President OR Election 2016 OR Republican OR Vote OR Campaign OR cruzcrew OR cruzers OR politics OR debate OR poll OR delegate OR voters OR campaign OR caucus OR candidate OR Presidential OR TedCruz

 Bernie Sanders: Bernie OR Sanders OR President OR Election 2016 OR Democrat OR Vote OR Campaign OR feelthebern OR bernie OR politics OR debate OR poll OR delegate OR voters OR campaign OR caucus OR candidate OR Presidential OR BernieSanders

 These type of keywords were used to collect the data for each city.

 Also we removed all emojis by running a program so we can get an accurate result by running the algorithm.

 After data was collected we separated the

data into positive and negative tweets based on sentiment analysis. This was done by running a program based on sentiment analysis where if the threshold value is 0.1 then its positive tweet else it’s a negative tweet.

 For example if the tweet contains the keywords such as happy or love then the tweet the regarded as a positive tweet.

 If the twee contains the keywords such as sad or hate then the tweet is considered as a negative tweet.

 After the data was separated into positive and negative tweets, training and testing data was formed for each city.

 Training data was formed as follows:

For example for New York City

 First we manually labelled the tweets as 0 and 1(where 0 is negative and 1 is positive) for one candidate which becomes our training data.

 So training data was formed which consists of tweet id, label and tweet text.

 Then the testing data was formed which included data of the other three candidates.

 Testing data consisted of tweet id and the tweet text.

 After that on the basis of these labels of tweet we ran the following algorithms to predict the number of positive tweets for each of the three candidates:

 SVM

 LR

 Naïve Bayes

 Decision Tree

 Association Rule Mining

 The candidate which has the most positive tweets has won in that region.

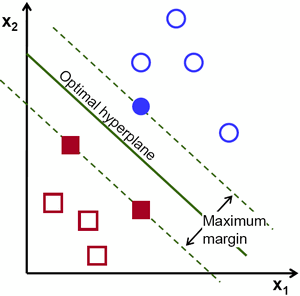
 **Algorithms**

 **SVM**

 It is also known as Support Vector

Machine.

 *A* ***Support Vector Machine*** *(****SVM****) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.*



- *Reference from* [*http://docs.opencv.org/2.4/*](http://docs.opencv.org/2.4/)*doc/tutoria ls/ml/introduction\_to\_svm/introductio n\_to\_svm.html*

 This algorithm was found to be the most accurate of all the algorithms used.

 For this algorithm we took our training data as our input and predicted the number of positive and negative outcomes.

 For example for New York City we took

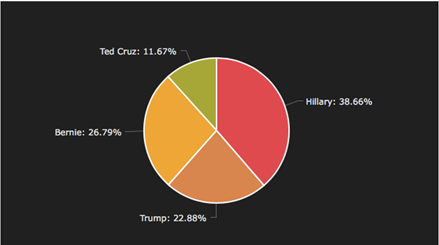
the training data of Hillary Clinton which helped us to predict the outcome of our testing data i.e. data for the other three candidates: Donald Trump, Bernie Sanders and Ted Cruz.

 From our results we found that Hillary

Clinton was the most popular in New York

City.

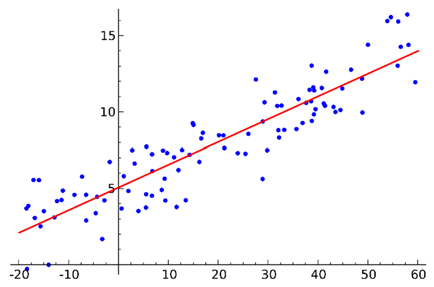
 Similarly for the other cities we calculated the most popular and the least popular among all candidates.



Example of pie chart for New York City using SVM

Logistics Regression

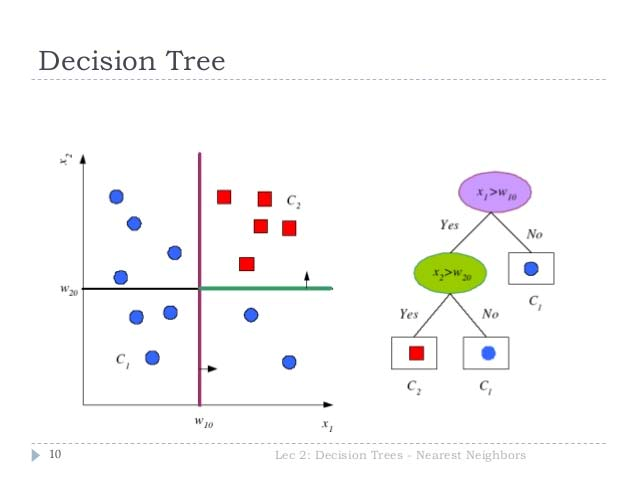
 *Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.*



***-****Reference from https://en.wikipedia.org/wiki/Logistic\_reg*

*ression*

Similar to SVM, for Logistic Regression also we used training data as our input and predicted the number of positive and negative outcomes.



 From our results we found that Hillary Clinton was the most popular in New York City whereas Ted Cruz was the least popular in New York.

**Naïve Bayes**

*Naive Bayes classifiers are a family of simple*

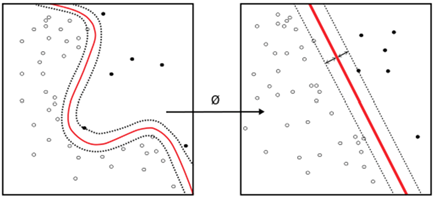
*probabilistic classifiers based on applying Bayes'*

*theorem with strong (naive) independence*

*assumptions between the features*.

*-Reference from https://en.wikipedia.org/wiki/Naive\_Bayes\_classifie*

*r*



Similar to Logistic Regression and SVM, we also used training data and testing data as our input and predicted the number of positive and negative outcomes.

 From our results we found that Hillary Clinton was the most popular in New York City whereas Ted Cruz was the least popular in New York.

**Decision Tree**

*Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.*

*-Reference from https://en.wikipedia.org/wiki/Decision\_tree\_learnin g*

*-Reference from* [*http://image.slidesharecdn.com/lecture02ml4ltmari*](http://image.slidesharecdn.com/lecture02ml4ltmari)

*nasantini2013-130827052029-*

*phpapp02/95/lecture-02-machine-learning-for- language-technology-decision-trees-and-nearest- neighbors-10-638.jpg?cb=1378716784*

 We frequently changed the training and testing using 10 folder cross validation to get the best training and testing data for our input.

This training data and testing data used as our input was used to predict the number of positive and negative outcomes.

 From our results we found that Hillary Clinton was the most popular in New York City whereas Ted Cruz was the least popular in New York.

**Association Rule Mining**

*Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.*

- *Reference from https://en.wikipedia.org/wiki/Associat ion\_rule\_learning*

We also implemented association rule mining in our project.

We took a dataset which contained the data of all the four candidates. From that data we found the support and confidence for each candidate and on the basis of that we got the outcome of our prediction for each city.

**Proposed Approaches**

We used the following methods for determining the best output.

 SVM

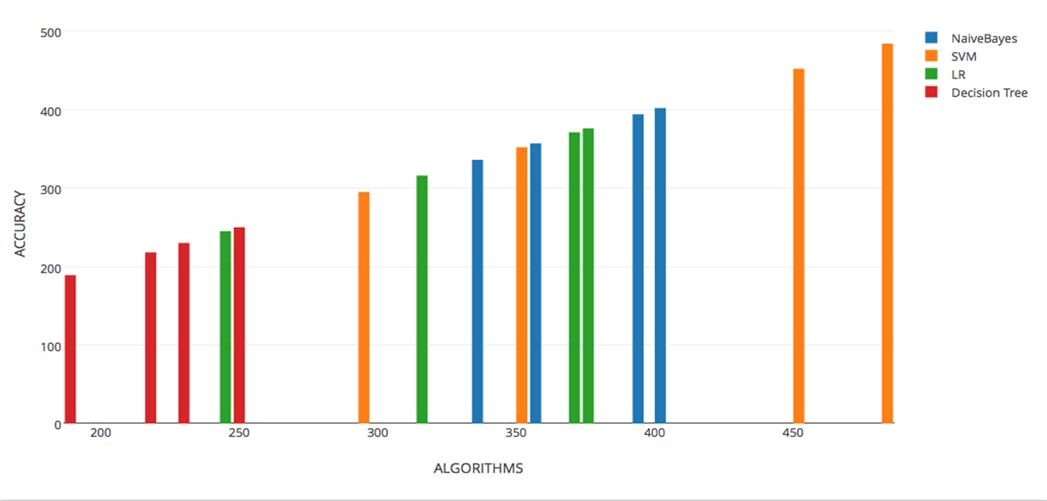
 Logistic Regression

 Decision Tree

 Naïve Bayes

 Association Rule Mining

Out of all these algorithms SVM was found to be the most accurate. This can be found out from the graph below:



**This is the figure we used in our project presentation to compare the accuracy of all the algorithms. Here SVM has the highest accuracy**

**System Design and Implementation**

The following tools and algorithms were used for implementing this project.

 Data Collection and labelling (Nisarg)

 Separation of emojis (Julia)

 Sentiment Analysis (Nisarg and Julia)

 API recall (Julia)

 Support Vector Machine (Nisarg, Neel and Julia)

 Logistic Regression (Julia)

 Naïve Bayes (Neel)

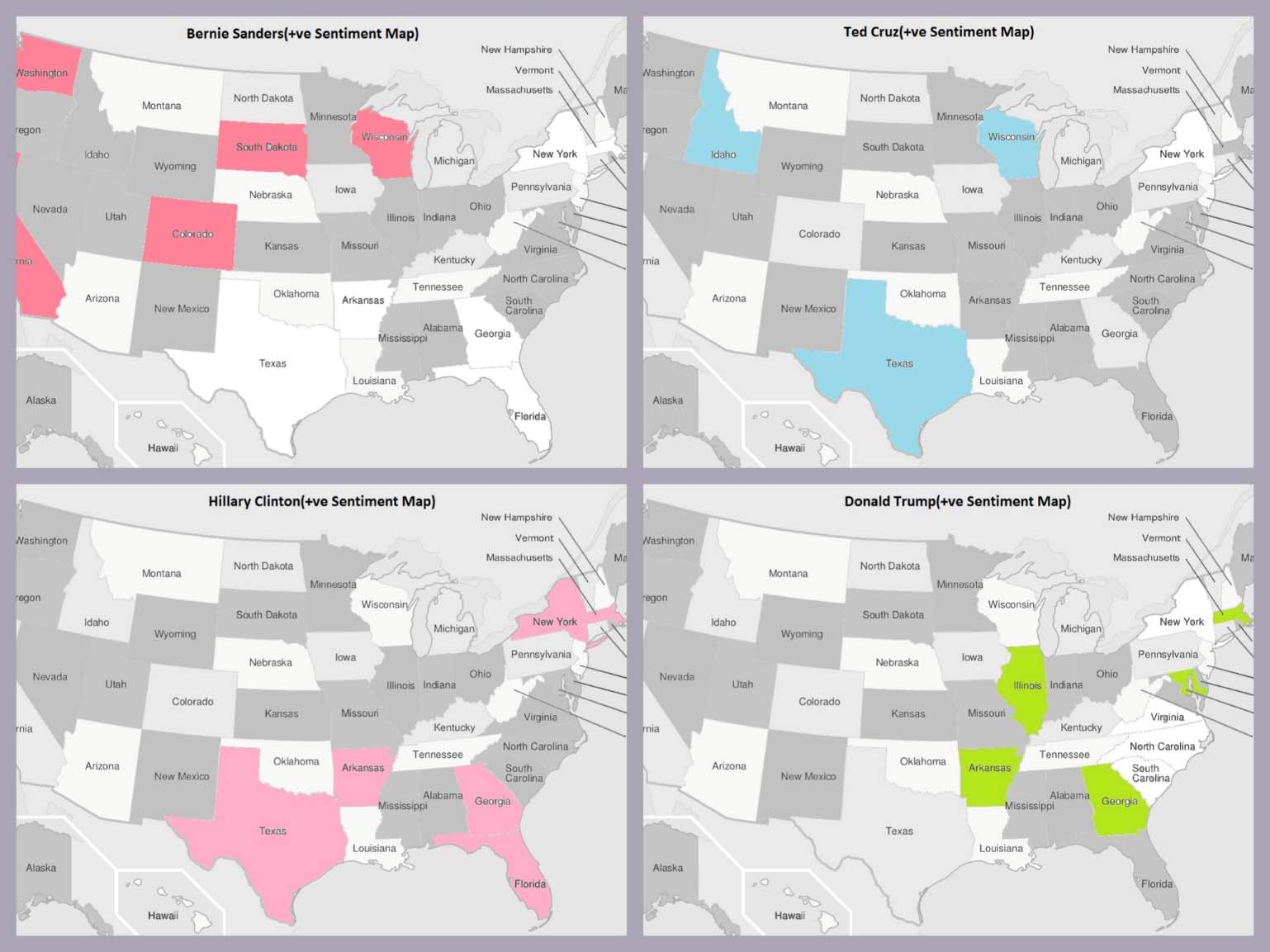
 Decision Tree ( Nisarg and Neel)

 Cross Validation (Neel)

 Association Rule Mining( Nisarg and

Neel)

 Data Visualization(Nisarg and Julia)



**This is the figure we used in our project presentation to show the states in which the candidates are most popular**

**Outcome of Project**

 Evaluated collected tweets to determine the least and most popular candidates

**Jacksonville** Hillary

|  |  |  |
| --- | --- | --- |
| Little Rock | Hillary  Clinton | Bernie  Sanders |
| Los Angeles | Bernie  Sanders | Donald  Trump |
| Louisville | Donald  Trump | Bernie  Sanders |
| Milwaukee | Ted Cruz | Donald  Trump |
| New York | Hillary  Clinton | Ted Cruz |
| Seattle | Bernie  Sanders | Hillary  Clinton |
| Sioux Falls | Hillary  Clinton | Ted Cruz |
| Washington  DC | Donald  Trump | Ted Cruz |

Clinton

Ted Cruz

|  |  |  |
| --- | --- | --- |
| **States** | **Most popular** | **Least popular** |
| Atlanta | Hillary  Clinton | Ted Cruz |
| Boise | Ted Cruz | Bernie  Sanders |
| Boston | Donald  Trump | Ted Cruz |
| Chicago | Donald  Trump | Bernie  Sanders |
| Denver | Bernie  Sanders | Ted Cruz |
| Houston | Hillary  Clinton | Donald Trump |

References

- [*http://docs.opencv.org/2.4/*](http://docs.opencv.org/2.4/)*doc/tutoria ls/ml/introduction\_to\_svm/introductio n\_to\_svm.html*

- *https://en.wikipedia.org/wiki/Naive\_B*

*ayes\_classifier*

- *https://en.wikipedia.org/wiki/Associat ion\_rule\_learning*

- [*http://image.slidesharecdn.com/lectur*](http://image.slidesharecdn.com/lectur) *e02ml4ltmarinasantini2013-*

*130827052029-phpapp02/95/lecture-*

*02-machine-learning-for-language- technology-decision-trees-and- nearest-neighbors-10-*

*638.jpg?cb=1378716784*

- *https://en.wikipedia.org/wiki/Decision*

*\_tree\_learning*

- *https://en.wikipedia.org/wiki/Logistic*

*\_regression*

- *http*[*s://www.kaggle.com/forums/f/15/k*](http://www.kaggle.com/forums/f/15/k) *aggleforum/t/4217/three-data-sets- and-k-foldvalidation*

- [*http://stackoverflow.com/questions/20*](http://stackoverflow.com/questions/20)

*587452/write-a-program-that- computes-meanmedium-maximum- and-standard-derivation*

- *https://github.com/asaini/Apriori/blob*

*/master/apriori.py*

- *https://gist.github.com/mblondel/5867*

*53*

- *https://github.com/rmaestre/K-fold- crossvalidation*

-