Predicting U.S. Presidential Election through mining social media data. (Twitter)

Palash Gore

University of Bridgeport, Department of Computer Science

Abstract

Data mining is a term that refers to extraction of knowledge or information hidden in large volumes of raw data. The purpose of this project is to predict the popularity of a candidate for US presidential election, 2016 form each state using social media for a given time interval [t1, t2], where t1 is the set of tweets observed between the timestamp t2. Until recently, political parties used information that limited pursing or reaching out to the masses which restricted the scope of a widespread campaign. The outcome of the project will help political parties make proper decision and target the right audience. This project is making use of twitter API which introduce simple concepts to analyze data. It will emphasize on techniques and considerations for mining large amount of data that is posted on twitter in real time.

Introduction

Social Networking websites as we know have been widely used for expressing opinions, sentiments, messages in the public domain through the use of Internet based text, image messages. From a pool of social networking websites such as Facebook, Reddit, WhatsApp, Qzone, Tumblr, Instagram, etcetera. Twitter has been a rich source of attraction to several researchers in important domains like advertisement, network analysis, prediction of democratic electoral events, consumer brands, movie box office, stock market, popularity of celebrities. The reason behind this attraction is due to its inherent openness for public consumption also, tweets allows us to get real-time insights from people's opinions in a virtual space which make the information quite reliable. In this project real time tweets will be analyzed on the basis of sentiment analysis or opinion mining to find the popularity.

Evaluation

Data Collection:

Data is collected and stored in Data set A in real time as the tweets are posted. Tweets are collected on the basis of word sets. For example, word set for positive sentiment, negative sentiment and neutral words. Also, tweets with names of the candidates, hashtags relating to the presidential campaign. Twitter Streaming API and Python Scripting language is used to retrieve data.

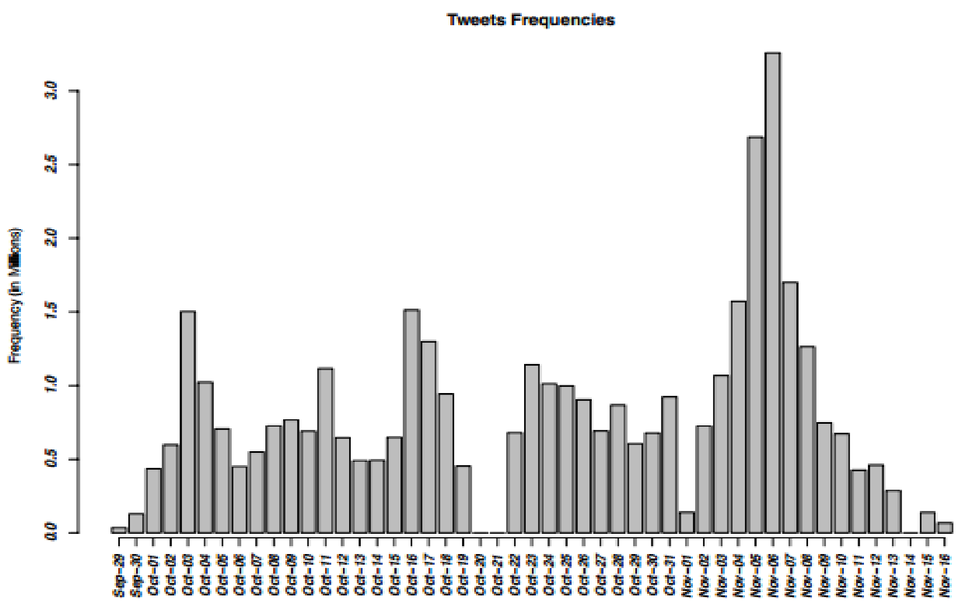
Normalizing and trimming data:

Tweets retrieved have data attached that may be useless. Hence, data is trimmed and only information to extract knowledge from is kept. For example, geographical location, time, date. Further, processed data is stored in Data set B for extracting knowledge and identifying patterns.

Data mining techniques:

Data is mined to extract patterns and useful information. Cluster analysis is performed to find the candidates popularity in a particular region. Further, mined information is used to construct histogram analysis. Clustering is essential when is useful when you want to group a small number of objects, may be cases or variables, depending on if you want to classify cases or examine relationships between the variables.

Model, Analysis and Results



Sentiment	Tweet
Positive	@sandrasays Obama gives a hope of bright future with his leadership skills.
Negative	Obama war policies affected the economy like hell
Neutral	@BarackObama to address Philadelphia tomorrow.

A typical tweet is basically a collection of text message and an image. All the contents listed below bundled together comprises of a 'tweet'.
Textual content: up to 140 characters that may include entities and Places.
Entities: essentially user mentions, hashtags, URL's, media.
Places: location in the real world

Histogram analysis is based on the frequency of tweets made in the timestamp t2. The algorithm for analysis will look for words that are inclined towards a specific sentiment.

The analysis is sentiment and opining mining based. An example from tweets made for the 2012 election shows how sentiments are categorized.

The result will be 3D printed based on popularity of the candidate his/her political party, weather he/she is a democrat or republican with state he/she is popular in.

Conclusion

Overall, my research on the topic reveals that social media data based behavior recognition analysis can increase the prediction accuracy along with sentimental analysis. Through analyzing social media data such as tweets, we can find interesting trends that can lead to better understanding, interpretation and insights However, it is essential to understand the parameters that influence the prediction in order to develop an efficient product with accuracy of results. Machine learning in most ways the best way to recognize patterns and predict results rather than taking human efforts on large sets of data.

Future Work

Given the nature of data found on social media there are some drawbacks and aspects that need to be considered while developing an algorithm that provides accurate results. Social media data is an important source of information for different types of content analysis. For a further study, more improved methods can be developed in mining social media that will contribute to a better understanding of the social media landscape in more details.