Who is US cheering for? 2016 election winner predicted using Twitter feed

Priyanka Dhingra
Rutgers University
Piscataway, NJ, USA
Email: priyanka.dhingra@.rutgers.edu

Feng Zeng
Rutgers University
Piscataway, NJ, USA
Email: fengzeng@scarletmail.rutgers.edu

Junjie Feng
Rutgers University
Piscataway, NJ, USA
Email: jf736@scarletmail.rutgers.edu

***Abstract*—** Using Sentiment Analysis to weight the polarity of Twitter reviews we can find who is positive, negative or neutral and can win the elections this time. This data from sentiment analysis can then be clustered for each of the presidential candidates to determine how many people are actually supporting them using k-means++ algorithm. This idea is derived from the non-thesis essay submission of Yelp: Rate my review by one of our team member.

I. Project Description

It will be the use of clustering algorithms and Classification algorithm that are inline with the proposed project types and sample algorithms on sakai. It is useful in numerous different perspectives. First of all the fun fact is that this determination affects everyone in US and World. The determination of the next president is probably the one big thing happening around the world privately or publicaly after the Environmental policies the world is working on. So this project might help immigrants make a plan ahead of time (in case Trump wins). There are further other questions that can be answered and I suppose media is using definitely such a technique to lure more viewers on their TV channels.

Considering feasiblity aspect, this project should be completed within the stipulated time because we have knowledge and a defined path. The only challenge that is forseen till now would be the collection of recent data. However, according to the explorations till now there are numerous API's available through twitter to make this happen.

It is a novel idea. It is a kind of Unsupervised learning approach that probably is not used anywhere. It is assumed that most problems like this will use a supervised learning approach. The clustering technique that will be employed defines that it is an unsupervised learning approach.

It is definitely a step by step process and the steps are clearly defined previously - first, collect data; second, run sentiment analysis; third, clustering; fourth, displaying results.

The project has four stages: Gathering, Design, Infrastructure Implementation, and User Interface.

A. Stage1 - The Requirement Gathering Stage.

This project is not restricted to the Twitter data only. This can use Facebook data as well. Different users with different mindset can use this. The methodology will remain the same. Consider a US presidential debate is broadcasted last night. Twitter will be full of tweets with oppositions and supports of one or the other presidential candidates. This data will be collected and can be used by media to draw some conclusions on how people are judging the candidates. This project can use different visualization techniques to actually draw mathematical results with genuine statistical conclusions.

Besides media the country itself can prepare for their next presidential candidates by looking at the graph calculated and the percentages on how bad and how good specific candidate will be using the front-end interface.

Also, listing the opinions of the people might affect other country voters and may or may not give a chance to other users who are not yet decisive about which presidential candidate is good. Therefore, on clicking the particular presidential candidate they can list all the reviews for that one candidate. To enable user specific interactions there can be login systems implemented however, this project seems to be for everyone and can accommodate features of using a facebook data or Twitter data in future and then logins can be used to determine the user's friends opinions. But since this is a topic for all and it should yield unbiased results the project is not recommending to use any user-specific results or visualizations. The deliverables for this stage include the following items

- A front end with dynamic clickable graphs and list of all the tweets if clicked
- As described above the users will not be specific individuals but industry-based so it can be media, voters, economists, immigrants or different countries. This can effect the economy of the country
- Some real world scenarios are described in the above discussion that makes the input and outputs well defined.
- Each of the team member will be equally responsible for carrying out tasks. Since it is a new horizon the project will require all three brains to work together. Still at the very micro management level the reports and presentation will be handled by one member by creating draft and then others will iterate and make amends. The testings and the development will be carried out together.

Deliverables for Stage1 is as follows:

- The general system description: The deliverables will include the sentiment analysis computed on the recent twitter data extracted.
- The three types of users (grouped by their data ac-

cess/update rights): The users types in here is - Economists, Media and Voters

- The user's interaction modes: The user's interaction modes will be visualizing data in different types of graphs by clicking on the types of available graph options. Looking at the tweets after they are clustered and hence can be defined as categorized tweets.
- The real world scenarios: Real world scenarios are, as follows.
  - Scenario1 description: Economists may want to visualize data using different visualizing methods to understand the most about the trending presidential candidate and then study about their policies.
  - Scenario2 description: Media can see presidential candidate's influence and can leverage from viewer's interest in the most talked about topic today by giving analysis drawn on factual data of tweets.
  - Scenario3 description: Voters, besides viewing can see the tweets as well after they are categorized on the conclusions drawn.
  - Scenario4 description: Economists are really an integral part for businesses as well. They can give real good inputs by this interface. Their analysis can also be added and they can be given a text editor to write their conclusions on each candidate.
  - Scenario5 description: Media can also view how the debate influenced public which topics of the debate are most discussed about by viewing tweets .
  - Scenario6 description: Voters can see which topics are actually important to judge a candidate according to this detailed and interactive report.
  - System Data Input for Scenarios: Whole dataset will be the Input
  - Input Data Types for Scenarios: Data type will be Strings with different polarity (good or bad)
  - System Data Output for Scenarios: Graph showing the trend
  - Output Data Types for Scenarios: Data type here will be weights given according to the polarity.

    Data input/output and data types for input/output will be same for all the scenarios in our case.
- Project Time line and Divison of Labor. The time line of the project will be based on the steps that will be performed and as mentioned they are- Data gathering, Sentiment analysis, Clustering and then visualizations. First three steps will be divided among three of us and the visualization, report and ppt are again three tasks that can be divided further equally.

References

[1] Roebuck, K. 1993 *Sentiment Analysis: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors.* 1993.

[2] G. Ganu, N. Elhadad, and A. Marian *Beyond the Stars: Improving Rating Predictions using Review Text Content.Twelfth International Workshop on the Web and Databases (WebDB 2009).* 2009

[3] N. Godbole, M. Srinivasaiah, and S. Skiena, *Large-Scale Sentiment Analysis for News and Blogs. ICWSM.* 2007

[4] P. D. Turney *humbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* 2001: Philadelphia, July 2002, pp. 417-424.