Comparison of Sentiment Analysis Approaches for Analyzing Short Texts

Marymount University

IT 489-A Caption Project

Dr. Thomas Narock

October 22, 2015

Honor Pledge: I acknowledge that the Capstone Project is an independent study project to be completed individually. On my honor, I have not received aid on my Capstone Project other than what was provided by my faculty mentor and any persons explicitly cited in my work. I further acknowledge that if I have given any aid to another student in this course, the instructor of this course was made aware of my contributions.

Table of Contents

I. Objective

In the past decade or two, the Internet has become a place where people can freely express their opinions and thoughts on just about anything there is on the Internet. These come in the forms of product reviews, blogs, and comments. The use of the Internet as a medium to convey ones opinion has increased in the past years as more and more people start using the Internet thanks to the ubiquity of smartphones. This is especially true in the case of social media websites where people can freely and easily use them to write, for example, Facebook statuses and Twitter tweets. According to a statistics by the International Telecommunication Union there are 3.2 billion Internet users globally (ITU, 2015). The biggest and the most popular social media sites are unarguably Facebook and Twitter. From recently released reports Facebook has over 1 billion users (Onfro, 2015), while Twitter has 316 million active users who create more than 500 million Tweets every day.

Data mining is a subfield of Computer Science concerned with revealing patterns and trends in large data sets, also known as Big Data. Consider a data set with 500 million tweets with attributes such as location of users, hashtags used, and user mentions. From this data set we can discover objective information such as topics that are trending in different countries. So, we can see what the people are talking about, but we do not know how they feel about it. If there is a single tweet, we can read it and understand the opinion of the person who wrote it; positive or negative. But we cannot go through all 500 million. This is the job for sentiment analysis. With sentiment analysis we can extract more subjective information from the data set, such as if the people are responding positively or negatively to the trending topic.

The objective of this project is to use different sentiment analysis tools and methods that use machine learning and lexicon-based approaches on real world Twitter data set and compare the

results of each tools and methods to see which ones best capture the sentiments. The theme of this data set will be the 2016 U.S. Presidential Elections. To accomplish the objective, the results will be compared to real life poll results of each candidate. Each election candidates and relevant tags or keywords (for example "#election2016") will be tracked on Twitter where any tweet that mentions these will be stored on a database. These data will then go through sentiment analysis and we can discover what kind of opinions, positive or negative, people have regarding each candidate. The results and the accuracy of the sentiment analysis tools and methods is not expected to be as accurate as approval ratings from election polls, however, the main goal is to discover a correlation between the results of the analysis and results of elections polls.

II. Client

The aim of this project is to compare various sentiment analysis tools and methods to see which one is most accurate. The data set will contain tweets about the 2016 U.S. Presidential Election. The presidential election candidates, and relevant hashtags and keywords will be tracked and stored in a database on a web server. Currently, as of October 22, 2015, there are 30 candidates from both Republican and Democratic parties that have registered for the election. Tweets mentioning these candidates are already being actively tracked and stored on a web server along with tweets that contain hashtags related to the election. There are a total of 75 hashtags and keywords that are being tracked.

III. Faculty Advisor

Dr. Michelle Liu has accepted to work with me as the faculty advisor and she expressed great interest in the project. She is an Associate Professor of Information Technology at Marymount University and currently teaches database technology, software engineering, and mobile

application courses. She got her doctoral degree from Boston University. I believe she will be of great help in this project because of her extensive knowledge and background in database technologies.

IV. Project Plan

Task Name	Duration	Start	Finish	Status
Project Topic Selection	15d	08/31/15	09/18/15	Complete
Set Up Server and Database	6d	09/18/15	09/25/15	Complete
Collect Data	41d	09/25/15	11/20/15	In Progress
Research Sentiment Analysis Tools and Methods	10d	10/12/15	10/23/15	In Progress
Prepare Project Draft	20d	09/25/15	10/22/15	In Progress
Choose Sentiment Analysis Tools and Methods	6d	10/23/15	10/30/15	Not Started
Start Sentiment Analysis on Collected Data	16d	10/30/15	11/20/15	Not Started
Organize and Visualize Results	10d	11/16/15	11/27/15	Not Started
Prepare Final Project Report	15d	11/16/15	12/04/15	Not Started

Table 1 – Project Tasks

Project tasks, dates, and durations can be found from the table above. The project can be divided into three parts. The first part of the project is focused on collecting the data and selecting the sentiment analysis tools and methods. In the second part, the focus is on doing the sentiment analysis. However, data will be continued to be collected until the end of the analysis. The third part is focused on cleaning, organizing the results and prepare them for the final report.

- Data Collection

A Linux server is set up that uses the "140dev Streaming API Framework" to connect to the Twitter Streaming API. The data received from Twitter is then stored in a MySQL database on the server.

- Data Analysis

  Metrics that will be used to determine the sentiment will be devised in this stage. Sentiment analysis will be carried out on the stored data using the chosen tools and methods. Each candidate will be given a score on a scale from 0 to 100, where higher the score the more positive the sentiment will be.

- Project Report

  In this last stage of the project, the results will be tidied up and visualized. These results will be then compared to election poll results and any correlation between the two will be studied. The final results will be put in the final project report.

V. Resources

There are number of hardware, software and other resources that are needed for this project, some free and some paid.

- Linux server – A fully managed Linux server is rented on Liquidweb web hosting services. The server has 8gb storage.

- Database Management System – Instead of using phpMyAdmin, MySQL Workbench is used to more efficiently and effectively manage and maintain data.

- Twitter API – The server will be connecting to the Twitter Public Streaming API services to stream tweet data from Twitter servers.

- 140dev Streaming API Framework – 140dev Streaming API Framework, a free PHP source code library, developed by Adam Green will be used to connect, collect, parse, and store the tweet data.
    - His work can be found on www.140dev.com.

VI. Project Details

Data collection

Each election candidates and relevant tags and keywords will be tracked on Twitter where any tweet that mentions these will be stored in the database. After the server is set up the tweet data is collected by connecting to the Twitter Streaming API using the framework provided by 140dev. Each tweet data from the Twitter servers comes in a raw form in JSON format. These raw tweets are stored on a cache table. A parser script is also continuously running that takes the raw tweets in JSON format and inserts it into these tables

- tweets – contains tweet ID number, tweet text, user ID number (unique IDs that is used by Twitter), screen name (@handle), user full name, tweet creation date and time, user location in longitude and latitude, and lastly a Boolean field that determines whether or not the tweet is a retweet of someone else's tweet.
- tweet_tags – contains tweet ID number, and tags (any hashtag that was included in the tweet text).
- tweet_urls – contains tweet ID number, and URLs (any URL that was included in the tweet text).

- tweet_mentions – tweet ID number, source user ID number, and target user ID number.
- users – contains user ID number, screen name, full name, location, user description (displayed on Twitter profile), account creation date, followers count, friends count, statuses count, time zone, and the last date the column was updated.

Candidates tracked by their user_id, which are unique user ID numbers that is used by Twitter. User ID numbers of each of the election candidates was obtained using free online services such as www.tweeterud.com, which looks up user_id based on their screen name or @handle. Only the official campaign Twitter accounts are tracked. Currently there are 30 accounts that are being tracked using their user_id. Furthermore, to increase the scope of the monitoring tag and keywords are also tracked. Currently there are 75 different tags and keywords that are being tracked. For example, tag and keywords include names of candidates, "#election2016", "#usaelection", "#GOP", "democrats", etc. When user_id is tracked, Twitter servers only return tweets that used the @handle of the user_id that we are tracking. For example, if we track Hillary Clinton's user_id, only the tweets of other users that contain the handle @HillaryClinton will be grabbed.

- "@HillaryClinton loved seeing you on Saturday night live! Funny!"

Therefore, by tracking the candidate's names as a keyword allows us to grab tweets cannot be tracked using the user_id.

- "Should Hillary Clinton Be the First Woman to Be Featured on the Ten Dollar Bill? http://t.co/mKeBdfYNW9"

As of October 22, 2015, the server on Liquidweb has collected about 8.9 million tweets which take up about 6.6Gib of space. Following are interesting simple figure from the tweets that have been collected so far.

user_id	screen_name	Full Name	Party	user_id mentions
1339835893	HillaryClinton	Hillary Clinton	Democratic	692279
25073877	realDonaldTrump	Donald Trump	Republican	624325
216776631	BernieSanders	Bernie Sanders	Democratic	432390
113047940	JebBush	Jeb Bush	Republican	136562
23022687	tedcruz	Ted Cruz	Republican	91867
1180379185	RealBenCarson	Ben Carson	Republican	37056
15416505	GovMikeHuckabee	Mike Huckabee	Republican	36839
15824288	MartinOMalley	Martin O'Malley	Democratic	34881
18020081	JohnKasich	John	Republican	32522
65691824	CarlyFiorina	Carly Fiorina	Republican	22957
15745368	marcorubio	Marco Rubio	Republican	21980
216881337	RandPaul	Rand Paul	Republican	16429
17078632	BobbyJindal	Bobby Jindal	Republican	11707
1347285918	ChrisChristie	Chris Christie	Republican	8364
58379000	ricksantorum	Rick Santorum	Republican	6980
432895323	LindseyGrahamSC	Lindsey Graham	Republican	4725
2865560724	GovernorPataki	George Pataki	Republican	1120
3021632183	gov_gilmore	Jim Gilmore	Republican	137

Table 2 – Number of user_id mentions of each candidates.

From the above table, we can see that the most talked about candidates are Hillary Clinton, Donald Trump, and Bernie Sanders.

Data Analysis

In this stage of the project, sentiment analysis will be done on the data using the chosen tools and methods. Upon sentiment analysis, each tweet about a certain candidate will be assigned a score from 0 to 100.

VII. Knowledge Being Applied

Knowledges that I have gained from my UNIX Operating System, Project Management, Database Technology, Statistics, and Decision Analysis courses are applicable in this project. My experience working as a Database Marketing Specialist is also extremely helpful because it made me familiar with database concepts and languages. Furthermore, the project requires me to learn new programming languages such as R language, and Python. In order to successfully complete this project extensive learning and research into data mining and sentiment analysis needs to be done.

VIII. Risk Factors

- Twitter Streaming API 1% rate limit

  One of the biggest problems in this project is that Twitter Streaming API gives a very limited access to tweets. Although it can track and store tweets in real time for free, only 1% of the total number tweets are passed on to the server. In order to gain wider access to more tweets or historical data one needs to use third party commercial products such as Gnip's Powertrack.

IX. Work Performed by Others

- 140dev Streaming API Framework – 140dev Streaming API Framework, a free PHP source code library, developed by Adam Green. It contains PHP scripts that allows the server to create a persistent connection to Twitter Streaming API, store raw tweets in the database, and a parser script that extract tweet information and puts them into their relevant tables. His work can be found on www.140dev.com

X. References

Bannister, K. (2015, January 26). Sentiment Analysis: What is it? Why use it? Retrieved 10 October 2015, from https://www.brandwatch.com/2015/01/understanding-sentiment-analysis/

Gallup, Inc. (2010). How Are Polls Conducted? Retrieved 3 October 2015, from https://onlinecourses.science.psu.edu/stat100/sites/onlinecourses.science.psu.edu.stat100/files/lesson04/How%20Are%20Polls%20Conducted%20FINAL.pdf

ITU. (2015). ICT Facts and Figures – The world in 2015. Geneva, Switzerland. Retrieved from http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

Medhat, W., Hassan, A., & Korashy, H. (1093). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. http://doi.org/10.1016/j.asej.2014.04.011

Narayanan, V., Arora, I., & Bhatia, A. (2013, September 16). Fast and accurate sentiment classification using an enhanced Naive Bayes model. http://doi.org/10.1007/978-3-642-41278-3_24

Onfro, J. D'. (2015). Facebook beats earnings expectations, but the stock sinks. Business Insider. Retrieved from http://www.businessinsider.com/facebook-q2-earnings-2015-7

Pang, B., & Lee, L. (2008, January 1). Opinion Mining and Sentiment Analysis. http://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Retrieved 1 October 2015, from http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Sentiment Analysis. (n.d.). Retrieved 3 October 2015, from https://semantria.com/sentiment-analysis

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011, January 6). Lexicon-based methods for sentiment analysis. http://doi.org/10.1162/COLI_a_00049