Twitter Live Feed Analysis and Storage in MongoDB

John Bellamy

12/17/2015

BDAA

Software Project

Professor Geoffrey Fox

**Abstract**

Sentiment analysis is a popular method that is gaining in popularity. Many sentiment analysis distributions can do a better job than humans at determining the sentiment of textual data. Perhaps the most popular use of sentiment analysis is for the analysis of social media. In this paper I will explore 36,000 tweets that contain the keywords “Clinton,” “Trump,” or both. I will then try to discover what the sentiment of these two individuals were over this period of time, as well as exploring what hashtags and at-symbols were most popular. Details of the techniques used are left to the accompanying iPython Notebook.

**Introduction**

Sentiment analysis is a method that can be used (especially with social media data) to determine how people feel about a topic. An effective sentiment program needs to not only take account for the polarity of a word, whether the word’s connotation is positive or negative, but also the semantics of the word, and the purpose of the word in the sentence. In this project, we will look at the sentiment of two chosen key-words: “Trump” and “Clinton.”

Using the public Twitter API I download just over 36,000 tweets over the period of December, 8th 2015 to December, 13th 2015 using the key-words listed above. Storing the tweets in a Mongo database, I then used an iPython Notebook to analyze the tweets.

**Textual Analysis of Popular Hashtags and At-Symbols**

Hashtags are a great way to allow others to track movements of causes and topics on social media. At-symbols, on Twitter, allow the person one is speaking of, and other Tweeters, to track mentions of their name, or a name of a person or thing one is tweeting about. It is for this reason I choose to analyze the hashtag— and at-symbol— usage in the corpus of tweets; the analysis of these symbols allow a glimpse into what is happening on Twitter at that time.

**Use of Parsimonious Rule-Based Sentiment Analysis**

Using a more advanced algorithm for sentiment analysis gives better results. Although polarity analysis is still important in techniques that account for semantics, an algorithm without semantics in sentiment analysis is much less accurate. The description from Vader’s program designers is below:

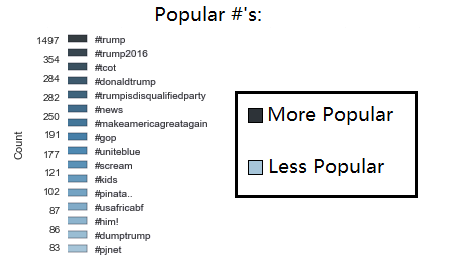
“*VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.”*

**Experiment Results**

The sentiments that come to any American’s mind when either “Clinton” or “Trump” is mentioned, might seem to obviate outcomes. However, the results that follow are not completely obvious and uninteresting. It is striking to first mention that tweets that contain the word “Trump” outnumber the tweets that contain either just “Clinton,” or “Trump” *and* “Clinton” by a factor of ten. Although this is not a political experiment, I cannot help but wonder if this fact is good for Clinton, as the Democratic front-runner. While it must be maintained that the purpose of any social media experiment is not generalization, it *does* appear that Trump was a much more popular topic during this time than Clinton.

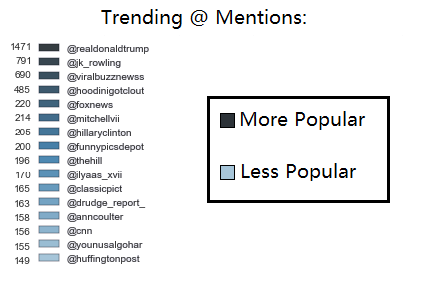
**Popular Hashtags**

Not surprisingly, “trump,” both with a capital and lower-case *T,* is the most popular hashtag. Perhaps more surprising is the fact that “trump2016” is the second most popular hashtag. This is surprising, because this seems to be supportive of Donald Trump’s candidacy, but belies the fact that Trump’s overall sentiment during this period was very negative overall, as we will see later. The 3rd most popular hashtag is “cot,” which is an acronym for “Conservatives on Twitter.” Finally, the 5th most popular hashtag was “trumpisdisqualifiedparty,” which is the hashtag in response to the White House Press Secretary’s statement that Trump’s comment about keeping Muslims out of America disqualifies him from the presidency. Below is a graph showing the most popular sixteen hashtags.



**Popular At-Symbols**

Again, not surprisingly, we see that the most popular at-mention was “realdonaldtrump,” which is Donald Trump’s official Twitter account. Also, not surprisingly we see that the author JK Rowling was a popular mention, since she had referred to Donald Trump as Voldemort. Amusingly, we can also look back at the hashtags and see that #14 is “him!,” which is an allusion to Voldemort as “him,” in the *Harry Potter* books and thus people on Twitter referring to Trump as “him;” read: Voldemort. Other interesting at-mentions are “hillaryclinton,” “foxnews” and “huffingtonpost.” Below is a graph showing the sixteen most popular at-mentions.



**Sentiment Analysis**

Reading through the remarkable research done by the authors of the VADER sentiment analysis software, we can see that VADER has an accuracy of 88%. This is remarkable, because as the software’s authors point out, human beings reviewing tweets are only about 84% accurate. While I will not go into detail or measuring accuracy here, it is worth noting the remarkable results VADER seems to produce.

After putting tweets into different lists, depending on whether the tweet contained the string “Trump,” or “Clinton,” or both, VADER was used to analyze the sentiment of each tweet in each list. The sentiment was given a score for: “pos,” “neg,” ”neu,” and also a compound score, which accounts for all three measures. The scores for each category were then summed, and a composite score calculated. The results and other details are in the chart below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Contains: | Positive Score | Negative Score | Neutral Score | Compound Score | Composite Score (Positive- Negative) |
| “Trump” | 2212.31 | 2608.50 | 23794.35 | -1309.10 | -396.19 |
| “Clinton” | 105.69 | 110.63 | 1211.68 | -9.78 | -4.94 |
| “Clinton” and “Trump” | 75.93 | 96.28 | 636.84 | -55.14 | -20.35 |

At first blush, it appears that Trump has a much more negative sentiment than Clinton. However, while apparently true, the sample size for the Tweets that contained *just* “Clinton” was 1,428. The sample size for the tweets that contained *just* “Trump” was 28,615. This means that the “Trump” list is approximately 20.04 times larger. Thus, if we “weight” the scores in the “Trump” list, the composite score falls to and the compound score falls to. This is assuming the sentiment in both cases are randomly-distributed, of course. Thus, either way you cut it, Trump seems to have a much more negative sentiment than Hillary Clinton, at the time this sample was taken, at least in these tweets.

Perhaps the most interesting, but not surprising, result is the fact that you are more likely to classify a comment as negative, using the composite score if it contains *both* “Trump” *and* “Clinton.” This result can be surmised to follow from Trump’s apparent polarity at the time of the sample, and the “lightning-rod” that Hillary Clinton is for Trump supporters. Or perhaps the fact that both candidates are polarizing.

**Conclusion**

While the fact remains that remarkable projects must follow from less-remarkable ones, this project shall remain the latter. The intention of this project was not to be remarkable, but to learn; mission accomplished. Through this semester, and for a few months before, I have learned a great deal about Python and the various tools available in this amazing programming language. I believe my Python skills are now at an intermediate level, and I will be writing more complicated programs and “graduate” to using my own Python user-defined classes very soon. In fact, much of what I have done here I may try to work into a program for my own use as a data scientist.

While it is difficult and dangerous to generalize, it can be said that Donald Trump had quite a negative sentiment during the period of time sampled. However, it is also clear, from the popularity of hashtags like “Trump2016” and “COT” that Donald Trump has his supporters.

My biggest regret for this project is that I did not make it to the classifier. I really wanted to take the results, from sentiment analysis and then build a Naïve Bayes classifier. But, time and ability are just not nearly there. However, this will be something I will try to finish over the break. Another regret is that I could not get as many tweets as I wanted.

Despite the flaws in the project, I hope that the project was at least entertaining, if not interesting. It was interesting and entertaining to create.

#Hillary2016.

Works Cited:

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.