How We React

Measuring Public Sentiment Toward Political Events and Issues Through Online Political Forums

Devin Cintron
Data-Driven Politics — Stanford University

June 11, 2018

1 Introduction

The use of online forums as a source for analysis is not a revolutionary concept in political science. Already, a number of researchers have made use of these rich datasets which have substantially expanded in both volume and particularity in recent years. Kropczynski, Cai, and Carroll [1], among many others, make direct analysis of political forums in an effort to determine the nature of the political deliberation that takes place — many concluding that issues arise in the realm of polarization and tenuous common ground. Still, some suggest that the anonymity granted by online discussions provides an avenue for "cross-cutting discussion" that rarely occurs in other mediums [2].

Regardless of the role these forums play in ideal formation and deliberation, it is clear that valuable insights can be made from direct observation. Klein, Clutton, and Polito [3], for instance, were able to analyze an online conspiracy forum to determine the overarching structure and individual characters that exist within conspiracy groups. The topic modeling they performed, in fact, was done on a conspiracy-focused category within the particular online forum website which this paper discusses: *Reddit.com*.

2 Data Collection

The dataset reviewed in this paper consists of approximately 10000 comments collected from two comment threads pulled from Reddit. Both threads were posted within hours of each other and consist of comments in response to a link to the article in which the Associated Press' confirmed Trump's victory in the 2016 presidential election. The first comment thread is pulled from /r/politics, a subreddit — or categorized forum — which is populated by left-leaning users discussing political issues. The second comment thread is pulled from /r/the_donald, a subreddit mostly populated by moderate conservatives, alt-right extremists, and other Trump-enthusiasts all gathered to discuss events surround Donald Trump's political life. Using Reddit's API, approximately 4000 of the top comments of the /r/politics thread and approximately 6000 of the /r/the_donald thread were collected. By the nature of the $comment - link - response$ structure of these threads, the data provides a valuable view into the reactions of these forums' participants to this event at event-time. There is an inherent limitation to the goal of sampling here in that the populations of the respective subreddits are not entirely liberal or entirely conservative due to the presence of periphery individuals with viewpoints inconsistent with that of the majority. Nevertheless, because of Reddit's up-voting mechanism in which majority opinion is essentially crowd sourced, these users' opinions are unlikely to be present in the top comments from which the sample is pulled.

3 Data Augmentations and Analysis

To demonstrate the richness of the data pulled from these forums, a number of augmentations and cleaning techniques were implemented:

1. **ASCII encoding of comments** such that special characters – many from usernames – do not cause issue when writing to file. Necessary to do if comments are to be stored either by writing to file or in any relational database

2. **Sentiment Analysis scoring** for individual comments and averages across threads

3. **Subjectivity Analysis** scoring for individual comments and averages across threads

4. **Activity-Motivating metric** – using part of speech tagging to track percentage of thread text which qualify as present tense verbs to have metric of comparison for 'call-to-action' language.

5. **Stemming Reduction, Common Word Filtering** filtered out insignificant words and performed stemming reduction on included words so as to be able to analyze the most common appearing word stems between the two threads.

3.1 ASCII Encoding of Comments (Data Cleaning)

Reddit comments very often contain special characters that will inevitably crash any attempts during the scraping process to store comments by writing them to an external file. Because of the presence of special characters (those outside common English characters, i.e. umlauts, etc.), UTF-8 encoding will result in many recurring 'junk' tokens which will invalidate attempts to track common word occurrences – ultimately junk tokens may appear to be among the most common words because they occur so frequently. More discussion is available in the comment section of the following source:

https://softwareengineering.stackexchange.com/questions/97247/what-is-the-advantage-of-choosing-ascii-encoding-over-utf-8.

```
file.write(top_level_comment.body.encode('ascii', 'ignore'))
```

3.2 Sentiment Analysis Scoring (engineering feature)

Using the Python library TextBlob, sentiment – polarity – scores were assigned to each comment. Scores ranged as floating points from -1.0 to +1.0, with an exact value of 0.0 being returned for failed analyses. Average polarity scores were taken from each comment thread for comparison:

Average for r/news: 0.297733404053
Average for r/the_donald: 0.420584254424

As one would predict, we see that the average sentiment for the conservative sampling, r/the_donald, is higher. However, one would suspect that the sentiment for r/news would be strictly negative. A possible explanation as to why it is non-negative is due to the fact that the corpus on which TextBlob's sentiment analysis is trained is a set of labeled movie reviews. If a superior labeled corpus more consistent with political commentary was available one might expect more accurate analysis.

3.3 Subjectivity Analysis Scoring (engineering feature)

Again using TextBlob, subjectivity scores were assigned to each comment. Scores ranged as floating points from 0.0 to +1.0. A score of +1 signifies a perfectly subjective token and 0.0 a perfectly objective token. Average subjectivity scores were taken from each comment thread for comparison:

Average for r/news: 0.549185415355

Average for r/the_donald: 0.59758310386

We see that both samples are moderately subjective on average, with a slightly more subjective average for the /the_donald sample.

3.4 Activity-Motivation Metric (engineering feature)

Part of speech tags were assigned to every word within the comment text of the two threads using the tagging feature of TextBlob trained on the Brown corpus. A count was maintained of the number of tags which fall under the category of present-tense verbs, i.e. ['VB', 'VBG', 'VBP', 'VBZ']

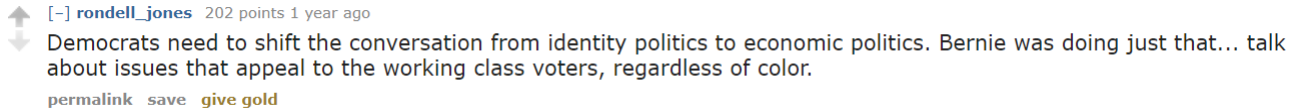
Using this count, a percentage score is assigned to each thread representing the amount of all words which constitute present tense verbs. This score is used as a rough metric as to the 'action' promoting nature of the thread, something of a rudimentary version of the *collective action events* referenced in King, Pan, and Roberts paper on modern government censorship in China [4].

The thread scores were as follows:

r/news: 0.150830501173

r/the_donald: 0.122936045576

Intuitively we would expect to see a higher score for the r/news thread as the more frustrated demographic collectively suggest actions going forward. A comment taken from the r/news thread serves as an example below:



3.5 Stemming Reduction, Common Word Filtering (cleaning)

In order to clean up the tokens taken from the two threads, two data cleaning techniques were carried out. First, a set of words were defined that would be filtered out (i.e. not included) when tracking the most common appearing words. The set was defined as:

.

[to,the,I,is,of,you,that,will,our,for,we,have,very,so,with

all,he,are,in,be,not,want,people,was,this,they,my,

know,who,I've,about,on,but,where,at,as,or,you,it,been,they're,the

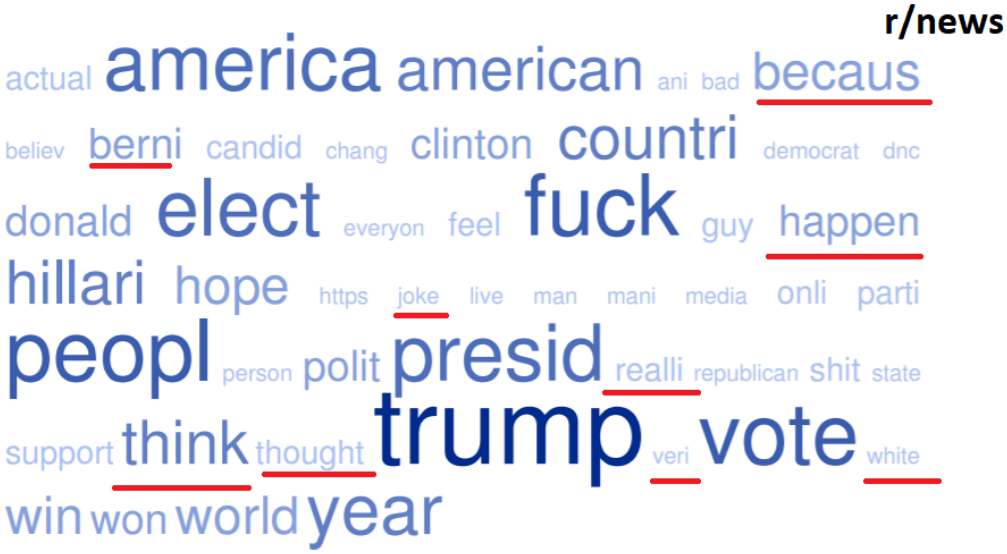
,usa,and,a,from,i'm,it's,what,did,an]

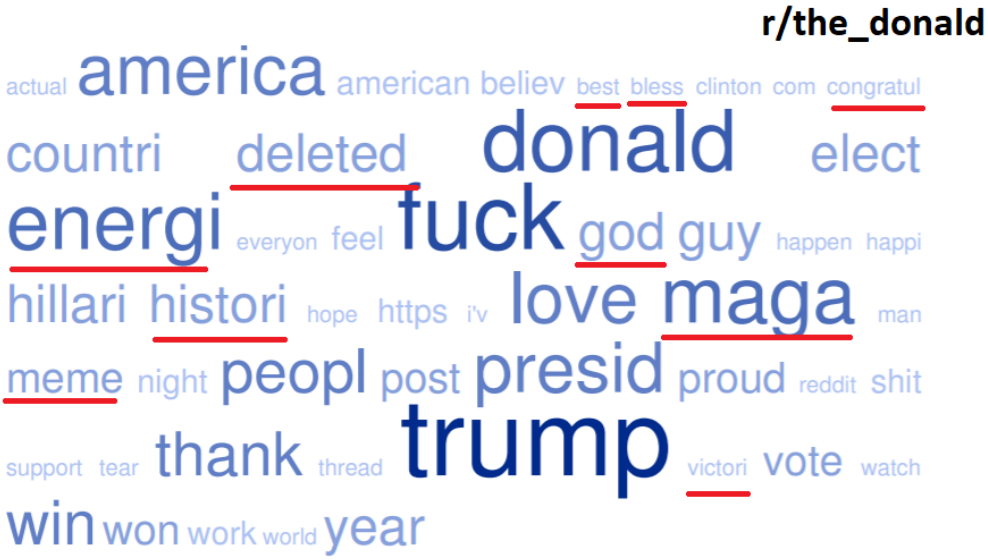
Additionally, stemming was carried out on each word in the comment threads using the Python NaturalLearningToolKit (NLTK) Snowball Stemmer. This stemmer is derived from Cambridge researcher Martin Porter's stemming algorithm introduced in the frequently cited 1980 paper *An algorithm for Suffix Stripping.* The necessity of carrying out stemming within these samples for analysis becomes immediately obvious when tracking the occurrences of expletives – once stemming is performed and words like f***, f***ing, and f***er are collected together we see that there are numerous expletives among the most common words within each thread.

4 Data Presentation and Applications

4.1 Word Frequencies WordCloud Comparison

After stemming and filtering out insignificant words, the following two word clouds were compiled independently for the two threads for the purpose of comparison:





What we see is a representation of the top 50 most common words – word stems, really – from the two threads wherein font size reflects frequency. It is interesting to note that many of the words are shared between the two. *Trump*, *presid (president)*, *f****, for instance, are amongst the very most common in both threads.

Looking instead at the disjunction, however, we can intuit some higher level general differences between the rhetoric and general sentiment of the two threads.

Some of the word stems unique to the r/news' most common 50 include "berni", "becaus", "happen", "think", and "reali". We see a common thread between these of words suggesting confusion and attempts to explain the event. For instance the word stems "berni" – from bernie – and "becaus" – from because – are suggestive of working to explain why the election ultimately went to Trump. The three other selected – "happen", "think", and "reali" – seem to suggest general shock and confusion, i.e. 'how could this have happened?' or 'is this reality?'.

Looking now to the word stems unique to r/the_donald's most common 50, we see a distinctly different narrative. A number of the word stems here suggest a general theme of celebration, including "bless", "best", "god", and "victori". There is additionally a number of stems which are symbolic of the general rhetoric of the online Trump camp including both "meme" and "maga". Finally, one particularly interesting word stem for r/the_donald is 'deleted'. The reason for the high frequency of 'deleted' is due to a large number of users electing to delete comments they previously made on the thread – or, beyond that, even deleting their entire account or having it deleted by Reddit's admin. It is unclear exactly what is behind this but an interesting potential explanation is a high number of bot comment submissions within the thread which were then either deleted by Reddit's admin or by the bots owner at a later time.

4.2 Issue-Targeted Group Sentiment

One of the particularly valuable benefits of using Reddit as a data source is the ease with which one can sample 'ideological' demographics. If one is curious about the opinions of libertarians on a subject they can sample r/libertarians, if curious about environmentalists they can sample r/green_party, etc. The array of political subreddits is incredibly broad, including everything from r/democrats to r/AnarchoPacifism to r/RonPaul. What this permits is direct sampling of nearly any particular ideological group of interest. It would, however, then be valuable to be able to sample that groups general sentiment toward a particular *issue*

of interest. For instance, one might be curious as to the comparing the general sentiment toward *abortion* between libertarians and moderate republicans.

We can carry out a reasonable bootstrapped implementation of this issue-targeted sentiment analysis using our 2016 Election response dataset. As a rudimentary example, we might be interested in comparing the general sentiment toward "Trump" between the conservative r/the_donald camp and the left-leaning r/news camp. Formally defining the process we will:

1. define some issue of interest i
2. loop through each comment within thread
3. if comment contains i, perform sentiment analysis on comment
4. add sentiment analysis score to running sum
5. return average sentiment score among comments which include i

This process provides us with a reasonable and computationally simple method for finding issue-specific sentiment scores within threads.

Carrying this out for the r/news and r/the_donald threads with $i = Trump$ we find the following scores:

Score for r/news for issue $trump = 0.0770538381658$

Score for r/the_donald for issue $trump = 0.249094380034$

There is a clearly much more favorable sentiment toward the issue of 'Trump' within the conservative r/the_donald sample. The non-negativity of the left-leaning r/news sample falls to the same issue of TextBlob's corpus being trained on non-political training data and the presence of sarcasm which is a consistent difficulty for NLP techniques regardless of training. Nevertheless, it is clear that even with these detriment that this is an effective process.

If we are interested in how these two samples perceive the status of the United States at election time we might query on the issue "America". Following the process with $i = America$ yields the following scores:

Score for r/news for issue $America = 0.0636411746802$

Score for r/the_donald for issue $America = 0.33834000261$

It is clear that those in the Trump camp appear to be commenting with a positive sentiment on the subject of America, though there may some noise due to the recurring phrase 'Make America Great Again' which is so prominent in any Trump-focused forum.

As a last query we may be interested in analyzing how these two samples feel toward 'God' at this point in time just after the election results came out:

Score for r/news for issue $God = 0.0383487914018$

Score for r/the_donald for issue $God = 0.234629021558$

It is apparent that God had a better standing among the conservatives at the time.

What is clear – and made evident by applying the technique to this dataset – is that valuable inferences can be made using this targeting approach. One can see how a more developed version of this process could possibly supplement, if not replace, traditional public opinion polls. Moreover, there are certain advantages to this process in that it is non-intrusive, easily ideologically-focused, and instantaneous.

4.3 Issue-Targeted Bag of Words

Just as one might be interested in analyzing the group sentiment toward a particular issue, it may also be valuable to analyze the specific *language* associated with the issue. So to say, it may be valuable to pull a bag of words related with some issue term i in order to analyze associations with that issue.

We can implement a simple, inexpensive process to compile this sample from our dataset. Formally, we:

1. define some issue of interest i
2. loop through each comment within thread
3. if comment contains i, collect word stems from the comment
4. return dictionary (map) of word stems with frequency count

Note: we additionally filter out word stems which come from common, insignificant words like *and*, *for*, *to*, etc.

In the context of this dataset an obvious term of interest is "Trump". If we carry out this bag of words collection process for the term $i = Trump$, the most frequent word stems in the respective bag of words for each thread are:

For r/news a bag of words of:

1. vote: 162
2. like: 117
3. if: 121
4. just: 117
5. go: 106

For r/the_donald a bag of words of:

1. presid: 125
2. donald: 100
3. fuck: 94
4. thank: 72
5. take: 67

The ability to compare these $BOWs$ for the respective threads gives rise to a couple of interesting observations from these most common word stems alone. It's clear that there is much discussion in the r/news camp surrounding voting, and in the r/the_donald sample many references to Trump with both his first name and the title president, neither of which are present in the most frequent stems for r/news. Presumably, the expletive here is used in celebration.

We may instead sample for a less thread-specific term. If we now carry out the collection process for the term $i = establishment$, the most frequent word stems in the respective bag of words for each thread are:

For r/news a bag of words including:

1. trump: 31
2. becaus: 16
3. like: 15
4. don't: 15
5. up: 14

For r/the_donald a bag of words including:

1. trump: 15
2. won: 8
3. against: 7
4. beat: 7
5. fight: 5

It is clear that for this query that the conservative sample includes a heavy anti-establishment rhetoric, seemingly celebrating what is perceived as a victory in an ongoing conflict, hence the word stems 'fight', 'beat', and 'against'.

In general we see that employing this process on these comment threads provides us a valuable means of collecting a set of issue-associated terms and language. Combined with the issue-based sentiment analysis of section 4.2, we have a simple and effective process for collecting group sentiment and associated language surrounding any particular issue. With the granularity provided by Reddit's subreddits, we are able to analyze the group sentiment and associated language for nearly any political event or political issue for nearly any political ideology.

5 Conclusion

As political issues and events become evermore frequent and public reactions evermore visceral, a low-cost and simple method for measuring these reactions becomes increasingly valuable. The ability to sample reactions a specific school shootings among libertarians or to police brutality among Neo-Progressives – and to do so in real-time – is now within the grasp of any researcher with an Internet connection. That reasonably valuable insights were able to be made from this paper's focused dataset using relatively rudimentary methods points to the richness of these Reddit comment threads as a source for analysis. Just as Facebook, Twitter, and many other social media sites continue to have a greater and greater relevance in political science research, we should expect Reddit to follow a similar pattern. In a political climate where issues and events are at a sprint, researchers have the duty to use the tools available to match their pace.

References

[1] Kropczynski, Cai, and Carroll, *Information Polity: The International Journal of Government & Democracy in the Information Age.* 2015, Vol. 20 Issue 2/3, p151-165.

[2] Kerill, *Information Polity: The International Journal of Government & Democracy in the Information Age.* 2009, Vol. 14 Issue 3, p219-232.

[3] Klein, Clutton, and Polito, *Frontiers in Psychology* 2018, 9 12p.

[4] King, Pan, and Roberts, *American Political Science Review* 2013, 107