Data Mining Reddit Comments

Sarah Barili
Sarah.Barili@colorado.edu

Kathryn Gray
Kathryn.Gray@colorado.edu

Roy DeJesus
Roy.DeJesus@colorado.edu

Jia Lin
Jia.Lin@colorado.edu

1. ABSTRACT

Reddit is a major social platform, where users can connect on a host of different topics. A data set containing the entire body of comments from the month of January 2017 has been published on the Reddit website for public use. This project is the exploration of the data set, more specifically, user activity in the 'politics' subreddit. The aim of the analyses scaled a broad range of topics, from using user sentiment to identify geographic location, to mining for patterns that could identify automated fake accounts. The data set was first filtered and stored using various processing techniques and warehousing techniques. After an exploration of the data in the r/politics subreddit, analyses could conclusively identify the most influential users, posting trends, and common topics of discussion. This project also offers some analyses via text mining of the actual comment bodies, and identifies some correlation between specific users, the topics they discussed the most, and their possible opinions and political leanings.

2. INTRODUCTION

Reddit is considered "The Front Page of the Internet by its users. It is driven by a strong, user-centered subculture that revolves around alternative knowledge and information discovery. As such, it can be sourced as a constant stream of new information regarding a vast range of topics. There is significant potential to derive meaningful patterns and trends from this data, and apply it to problems concerning anything from natural disaster prediction to mental health diagnosis. The main difference between this media platform and others is its highly unregulated nature; it could potentially provide a more transparent insight into the sentiment of its population of users. In turn, this could be used to identify majority opinions and social trends that are not necessarily publicized.

The goal of this project is to find meaningful patterns in Reddit user comment data, specifically comments from the r/politics subreddit. We address several of the following challenges regarding users and comments in the r/politics subreddit:

- Determining which words and phrases were used most in r/politics
- Cross checking users posting in the r/politics subreddit with other comments and frequently used subreddits
- Identifying the most influential users in the r/politics
- Determine the comment distribution through the month of January
- Identifying user sentiment on specific political topics

These analyses will help determine whether political leanings or other knowledge can be identified from Reddit comment data. This is a relevant issue because the recent election polls did not accurately predict results. While we are not planning to look at multiple months, data mining techniques like this could be used in the future to discern whether using online forums is a more accurate way to determine voting results. We could also use this to determine user ideas about current events, such as the current policies, which could be more accurate or informed than polls.

3. LITERATURE SURVEY

Many research projects followed from the publication of Reddit comment data. There has been work done on analyzing users, types of comments, linguistics, and comments on specific topics.

3.1 Reddit and Behavioral Therapy

In "Role of the Social Web in Behavioral Therapy researchers analyzed mental illnesses through the user's comments[3]. This was accomplished in a multi-step process. First, the language of "self-disclosure among Reddit users was identified with text mining. Researchers mined the frequent words in three identified mental health-oriented subReddits, and compared these words to a psycholinguistic lexicon to determine which types of words appeared the most often (i.e. words that are 'emotionally expressive'). Using this, they built a statistical model to predict the factors that determine maximum social support on posts in these subReddits. Finally, they used the information found in these subreddits to identify the role of "disinhibition in the Reddit community due to analyze its effectiveness in getting users to disclose their mental illness. In this way, Reddit posts could be proven useful to predict or identify mental illness in its users as a means to bring better assistance to these users.

3.2 Reddit Scoring and User Analysis

There has also been research into the types of comments that users post. Researchers studied which comments had the most upvotes, based on when they were posted and how long the user had been on Reddit [5]. These researchers also had research on how many comments a user had posted based on how long they had been a part of the Reddit

community. They also researched how many upvotes and downvotes a comment got based both on how many upvotes and downvotes its parent got and how high in the comment tree the comment was.

3.3 Analysis of German Subreddits

In "German Reddit Corpus, researcher Adrien Barbaresi analyzed comment contents to build a linguistic corpus of German language based on its use in German sub-Reddits[2]. The data was visualized to map the possible geographic locations of German-speaking users. He first utilized language detection software to store words and phrases as tokens in a dictionary structure. From this he was able to derive useful information, such as the frequency of German word usage in sub-Reddits, and the most common nicknames used by authors in these posts. By comparing the dictionary of tokens to a lexicon of nouns that describes cities, and a geonames database that maps cities to specific locations, he was able to geographically locate the posts based on city name mentioned in comments.

3.4 Analysis of Election Text Data

In "Analyzing Trump v. Clinton text data at Reddit, Emily Gao analyzed the comments from the r/politics sub-Reddit for July 2016 [4]. By identifying the most popular words and phrases in these comments, and finding correlation between the timestamp of the most frequent words used and significant happenings in the news, Gao was able to identify what Reddit users were most concerned about in the months preceding the 2016 elections. First, she mined the data set for the most frequently used tokens used in the sub-Reddit in July 2017. Second, she analyzed the frequency of each word according to the timestamp of the comment or post, and provided a useful visualization to identify certain periods during the month that these words seemed to spike in popularity. Using this data, she then identified the top 20 most popular comments (based on votes) corresponding to these days, and derived a collective user sentiment on these topics and words. For example, The word "Hillary, and the topic "email servers seemed to spike in popularity on July 5th, and mining the top comments for this day revealed collective user sentiment that could have been used to more accurately predict the results of the 2016 election that followed several months later.

4. DATA SET

The data set being used was found on the Reddit website[1], published by Reddit user StuckintheMatrix. It is approximately 42GB of user comments from the month of January 2017. StuckintheMatrix created this data set by making repeated, automated calls to Reddit's Application Program Interface, API, to retrieve the user activity data. The data is stored in JavaScript Object Notation, or JSON, string format; each object represents a comment. Each field in the JSON object can be compared to a tuple in a non-relational database, where the fields of the JSON object stand for the attributes of the database. An example 'node' in the data is pictured below:

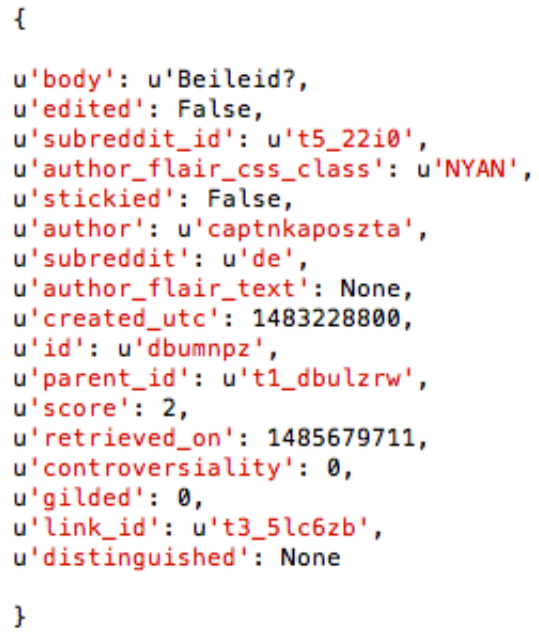

Figure 1: An example JSON 'node' from the data set.

There are 78,946,585 comments in the data set and each object has the following fields: Body, Edited, Author, subreddit, subreddit ID, Author Flair, Stickied, ID, Parent ID, Score, Retrieved On, Controversially, Link ID.

The body is the most important field utilized. It contains the actual text of the comments, which was is used to mine for frequent words and phrases. The edited field allows us to see if, at any point of the comment's existance, edits were ever made.

The subreddit and subreddit ID was be useful for our initial data reduction as well as determining the other subreddits where the top influencers also posted.

Author flair css class and flair text field displays the reddit user's css or text color flair details.

Stickied was a boolean field. A stickied comment refers to a comment that has been pinned to the top of a page. This is typically for used to specify rules of the subreddit or important and frequently used documents.

The Author field stores the username of the author of each comment which was crucial for tracking users to determine how the they might feel about a specific issue.

The score, which tracks the upvotes and downvotes, was used to determine the most and least popular comments of a given user.

The created on field contains the creation timestamp of the object in the Unix time system. Similarly the retrieved on field contains the Unix timestamp of when the comment was retrieved from Reddit's API.

The controversiality field is a interger field that only supports a 0 got non-controversial comments or a 1 for controversial comments.

Reddit supports donations to users for certain comments directly from other Reddit users. The gilded field displays how much gold a comment has received.

Link ids are used for direct url paths.

Distinguised comments are comments that have been awarded by moderators.

The most fields used in this analysis are the subreddit, body, author, score, and created on. The data was not missing any values, and did not have any incorrect data that might

have skewed our results. After looking it over, we found we did not have to take any steps to remove inconsistencies or missing values.

5. EVALUATION METHODS

Because of the broad range of information we sought to find, several different techniques were used for data preprocessing and warehousing. For numerical, timeseries, and user data, preprocessing and warehousing was used with the Python Pandas library. All methods were tested on on a sample consisting of the first partition of the r/politics comment data before being applied to the entire set to improve speed and efficiency during testing and development phases. To text mining the comment body fields, WEKA was used.

5.1 Tools

5.1.1 JSON

Our dataset is JavaScript Object Notation, or JSON, formatted. The JSON file data is simple and readable, comparable to a python dictionary. This was useful to us since JSON is easily read by many softwares, including python.

5.1.2 Python

To take advantage of the JSON format, we decided to use Python as our main language. The JSON data is parallel to the structure of a python dictionary, and python has a large and comprehensive standard library, which contains various convenient JSON oriented tools and methods. Python's json library was mostly sufficient for this project but frequent use of the simplejson library was also used. simplejson allowed for detailed error messages to be displayed which was used for debugging purposes. The textblob library was a great library for text processing. The Pandas Library contains various Python specific tools for data analysis and data display with the convenience of easy cooperation between the Python scripts.

5.1.3 Jupyter Notebook

Jupyter notebook was used with the python programming language to benchmark progress and visualize program output. Jupyter is a useful way to keep track of live code, data analysis and visualizations, and written reports in one place. It is an open source software project that functions with many programming languages. Jupyter Notebooks store progress linearly, so it was well suited for this project. Coupled with python's pandas library, Jupyter was a powerful tool for our group, and was used in the data mining process from Pre-processing to Visualization .

5.1.4 WEKA

The Waikato Environment for Knowledge Analysis, or WEKA, was used to count frequent words in the comment body field. WEKA is a Java based machine learning software, put out by the University of Waikato. It has a large collection of data mining and machine learning algorithms, which allows for quick access to many different techniques. The algorithms used for this project included text collection and organization of frequently used words and phrases. WEKA is also convenient for text data presentation. WEKA was also able to remove useless comments, sample the data to help runtime, and help gain a better understanding of the comments.

5.2 Numerical, timeseries, and user data

5.2.1 Data Preprocessing

Significant data reduction was required to conduct meaningful analysis. The r/politics subreddit consists of approximately 1.2 GB of data, and the remaining 42GB of the Reddit comment body was largely irrelevant to us; reducing to one subreddit significantly reduced the time spent on data mining the set, and eliminated unneeded data. The nodes in this data set were very consistent, so required cleaning was minimal. Python's JSON library was used to reduce the dataset, using numerosity reduction via sampling. Python's JSON library is simple and easy to use; we developed a program that read in our 42GB full data set, processed each JSON object node as a line, decoded and searched the fields to reduce based on subreddit name, and encode back into a JSON object to write to an outfile. The JSON library experiences complications when trying to encode a very large file as a JSON object (>1.5 MB). To accommodate this, the original data set was partitioned into 8 sample sizes, and another program was developed to move the JSON comment nodes from the large data set to a separate file in a python list data structure. The resulting reduced data set was stored as 8 smaller files of comment nodes stored in a list containing only data from the r/politics subreddit. Comments with author field as '[deleted]' were treated as invalid entries and removed from the dataset in this step as well.

5.2.2 Data Warehousing

The JSON objects in the original data file had uniform node attributes, and could easily be stored and operated on as a relational database structure. Upon further investigation into the data set, moving the original JSON objects into a parallel database structure was unnecessary. For data mining and analysis, the JSON data stored in python lists as dictionaries was directly operated on. For both the time series analysis and the identification of frequent users, a program was developed to iterate through a list of JSON objects, and store the relevant fields as a python list of tuples. Lists and arrays usually have poor performance time on large data sets, so these programs were tested on a sample of data before working with the entire r/politics subreddit. These lists were then passed through python methods discussed under Data Mining techniques; the results were stored in simple Pandas Dataframes to make graphing and visualization of results easier.

5.2.3 Data Mining Techniques

The method Counter, from the python library collections was then invoked on the list to count the occurrences of each author. Counter uses frequent pattern matching algorithms used on multisets and bags, similar to the apriori algorithm. Counter output is stored as a list of tuples (item, frequency), which was easily moved to a small dataframe, pictured below:

To conduct a time series analysis of the comment density, data transformation on each comment's creation time-

	Author	Comments
0	LeMot-Juste	2967
1	kiarra33	2149
2	takeashill_pill	1985
3	TinyBaron	1813
4	pizzashill	1779
5	AtomicKoala	1758
6	Donald_J_Putin	1751
7	ishabad	1728
8	DiscoConspiracy	1679
9	VROF	1661

Figure 2: Dataframe of Top Influencers, with author name and posting frequency listed below

stamp was needed. The field was normalized from UNIX timestamp to python datetime structure, formatted YYYY-MM-DD HH:MM:SS with milliseconds. A graph of this data revealed high noise, so the data was then discretized into 31 equal width bins of 24-hours each (neatly organized into days). The same Counter method was invoked on the time-stamp attribute after smoothing, to aggregate the daily comment density, stored in a dataframe with a similar format.

Further investigation of the frequent comment times led to searching the daily comment density for the top 5 influencers. First, the data was clustered by comment based on the field 'author' for each of the top five authors and was stored in a two-dimensional dataframe. A new dataframe was formed by performing a database style outer join on the individual user dataframes. Each author served as a field in the new dataframe, and the 24-hour time-stamp bins served as tuples. 'Nan' or NULL fields were normalized to 0, as a Nan value indicates user did not post to r/politics on this day.

	Le-Mot Juste	Kiarra33	takeashill_pill	DonaldPutin	DiscoConspiracy
Date					
2017-01-05	213	39	95.0	0.0	94.0
2017-01-04	188	59	75.0	0.0	120.0
2017-01-20	169	61	78.0	150.0	110.0
2017-01-08	127	20	83.0	38.0	30.0
2017-01-21	124	49	67.0	65.0	5.0

Figure 3: Dataframe of Comment Density of top 5 users

5.3 WEKA for Text Analysis

WEKA was used to analyze user and the complete r/politics texts. WEKA's functions on finding most common strings were used heavily to find the most common words and ngrams in different scenarios. For the complete r/politics, the comments were broken up by binning into scores. The bins included scores less than -20, scores between -20 and -1, scores equal to 0, scores between 1 and 20, and scores greater than 20. This helped break up the large amount of data into smaller subsets. Only the comments with negative scores were used to try and understand which topics people felt strongly against.

5.3.1 Methods Used on Compete Subreddit

WEKA's Resample function was used to get a smaller sample of the data. This resampling helped with runtime and some memory issues with WEKA. Both sampling without replacement and stratified sampling were tried, but both returned similar results. Once all the comments were found, StringToWordVector was used to find the top words. WEKA was also used to determine which words were the difference between low rated comments and comments who rated below -20. In this case, the Iterated Lovins Stemmer was used. The Lovins Stemmer works by truncating the word, removing the endings. It removes endings from words and has a wide range of endings it looks for. It will remove everything from "s to "alistically, so it was a useful method to get a word down to its base stem. This helped determine what users were most passionate about since there were so many words we could have ended up with trump/trump's as two separate words, but since we used a stemmer, they would fall into the same stem. The NGramTokenizer was also used. This meant that not only would single words be kept, but phrases up to five stems would be kept, "five-grams. Most of the words that came back were only unigrams and bigrams, especially when looking at the top few words. A couple of the top bigrams were "the dnc, "the gop, and "fake new, while some of the top unigrams were "trump and "lol. To determine which words would help classify a comment as receiving a negative score less than -20, WEKA's attribute selection was used. Attribute selection picked the best words to split on if making a decision tree, using info gain ratio.

5.3.2 Methods Used on Single Users

For the single users, a similar method was used with a few key differences. Sampling was not necessary since there were only a few thousand comments. There was no stemming or ngrams found because this muddled our data. For example, when using stemming and ngrams, phrases like "in the would come up for most common used phrases. This was not helpful towards the goal of determining what a specific user was talking about, so it was more useful to not use ngrams and a stemmer in this case. Stop words were also more important when looking at only single users. Stop words will remove words like "the, words that will not be useful to determine what users care about. In WEKA, the stopwords Rainbow function was used. Rainbow performs a statistical text classification that helps create the stopwords from the text. This was extremely useful in the case of specific users since it helped narrow the search field in a meaningful way, and we received results that both made sense and helped us make sense of a user's comments.

6. RESULTS

The results we achieved spanned from analyzing top user comments to determining when the most comments were posted. These included discovering the day most posted, in our data set this was Inauguration Day and President Obama's Farewell Address. We discovered that specific users had specific things they cared about, for instance, one user would care more about information, while others were more concerned with the party system.

6.1 Most Frequent Posters

The results of the analysis of the most frequent posters in r/politics revealed a high volume of automated moderator comments.

```
Number of comments:  2027483
Number of Authors:  156320
Top ten influencers:
('AutoModerator', 45654)
('PoliticsModeratorBot', 9085)
('LeMot-Juste', 2967)
('kiarra33', 2149)
('takeashill_pill', 1985)
('TinyBaron', 1813)
('pizzashill', 1779)
('AtomicKoala', 1758)
('Donald_J_Putin', 1751)
('ishabad', 1728)
```

```
Number of comments:  1972744
Number of Authors:  156318
Top ten influencers:
('LeMot-Juste', 2967)
('kiarra33', 2149)
('takeashill_pill', 1985)
('TinyBaron', 1813)
('pizzashill', 1779)
('AtomicKoala', 1758)
('Donald_J_Putin', 1751)
('ishabad', 1728)
('DiscoConspiracy', 1679)
('VROF', 1661)
```

Figure 4: Comparison of the program output for 'Top Influencers', with and without automated moderator posts

54,739 of the comments from the original 2,027,483 total comments from the politics dataset were found to be automated moderator comments, or 2.7 percent of the total comment volume. A second iteration was conducted on after the set was further reduced to remove as many moderator comments as possible, as the findings were significant enough to effect the program output and later attempts at text mining the user comments.

The identification of the most frequent posters in r/politics was used in several following analyses. The actual comments posted by the top users were extracted and moved to a comma separated values, or csv, file, from which useful information was gathered using the text mining tools in WEKA. WEKA was able to find a user's top and least rated comment as well as an average comment rating. WEKA first took all the comments in the csv file and split up words. It then returned the top word used by the users.

6.2 Most Common Words in r/politics

The first run through showed the most common words as the words used by moderators and bots. So we decided to look at the data with really high scores and really low scores. This would give us a better understanding of issues people care the most passionately about. The stems were first narrowed down by looking at the top 1000 words, and then these words were used to determine the best words to determine the difference between a low and high negative score. "Trump scored highest among these words (not including "[deleted] and "[removed] comments), followed by "lol, "republican, "fak, "gop, and "fake new. This shows that in January, many people were concerned about fake news and Trump. This makes sense with news stories coming out at the time, and nothing is really surprising in this analysis.

6.3 Specific Users' Commonly Used Words

When looking at specific users, we did not use a stemmer. We found that because there are a lot less words in just a single user's comments, the stemmer only served to hide what the user really cared about. One of the instances of this was the stem "don. When used as a stem, it is unclear whether this is referring to "donald or "don't. So the stemmer was not used on separate users' comments because we could gain more information without using it.

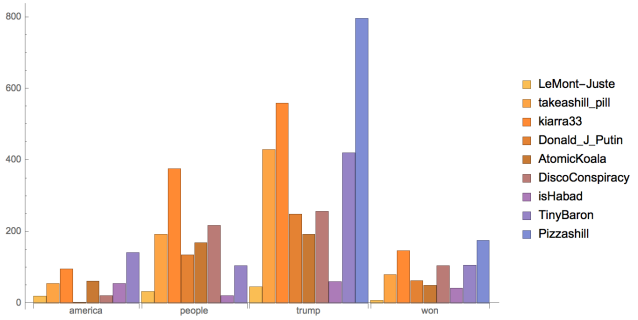

Figure 5: Frequency of Words by Top Commenters

One of the most interesting things from this analysis was seeing the top commenters had issues that they talked about more than others. This was interesting because just reading through a user's comments you may not be able to see the most important thing to them, but when looking at their comments as a whole, a new picture of the person posting emerged. They did not care about all hot topics equally. By looking at their top used words, it became clear which topic they cared deeply about. One user had a top words of "assange, "russia, and "news, this user also posted on subreddits such as r/hacking, showing that this user cared deeply about the spread of information. Another user's top words included "russia, "russian, "election and "free, perhaps showing this user cares about whether the election was influenced by Russia. Top users also did not use the same words equally. While some people used the word "america a lot, others barely used it at all. In fact, the top commenter overall had little overlap to other top commentors most frequently used words. This was surprising because even if it was not in a specific user's top words, it was expected that some of the other top words from other top commenters would be similar. However, we found that was not the case, and word use was much more specified for a given user.

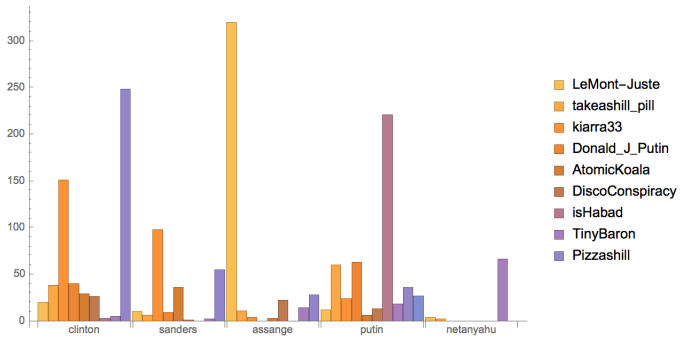

Figure 6: Frequency of Names used in Comments by Top Commenters

6.4 Specific Users' Scores

We looked at the top users scores as well, but did not find anything conclusive in this. Most of the scores were gathered in the -10 to 20 range, most lying between -5 and 5. There were a few outlier comments, ones that received scores in the thousands, and a few outliers that received scores in the negatives around -30 to -100. When looking at these comments, there was no real correlation we could draw between what these specific users said that would cause the comment to have a very good or very bad score. A typical user score is shown in the figure below, with the extremely high and extremely low scores taken out.

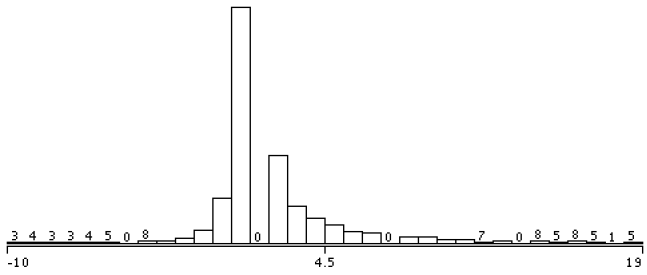

Figure 7: User Scores in the Average Score Range.

Average user scores were rarely above 10. Only one of the most frequent commenters had a score above 10, and while we did not analyze why this was the case, this user's top comment had 17 comments with scores above 1000, and only 23 comments below 0. The user only had one comment in a subreddit outside of r/politics, in r/technology. The user also did not have many top words that were not other users' top words. The two words that differed in this user's top words list were "white and "world. Neither of these really show why this user scored so highly on many of their comments. Further analysis would have to be done to determine a better reason.

6.5 Analysis of Other Posts by Users

Using WEKA, we were also able to look more closely at the individual behavior of these top influencers outside of r/politics. In several instances, useful information about the author was obtained. Figure 7 shows which other subreddits the user 'kiarra33' posted in. Included in this author's other posted subreddits are 'WayoftheBern' and 'SandersForPresident'; considering this and the author's most frequently used words, among which was "sanders, used 98 times, an intuitive correlation could be drawn between this users posting in other subreddits and their political affiliation as a Bernie Sanders supporter. When actually looking at this user's comments, however, we find a much more interesting relationship with their support of Sanders. They say "So I love Bernie Sanders but unless he can figure out on how to make a single player healthcare plan that can be 100% paid by federally, until then her[Clinton] plan better and more progressive. This user seems to take issue with several of Sanders' policies, something that is hard to see from a mere overview of their posts.

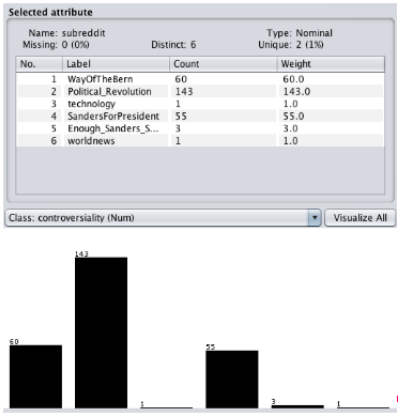

Figure 8: WEKA analysis of author 'kiarra33', including postings in other subreddits, and most frequently used words

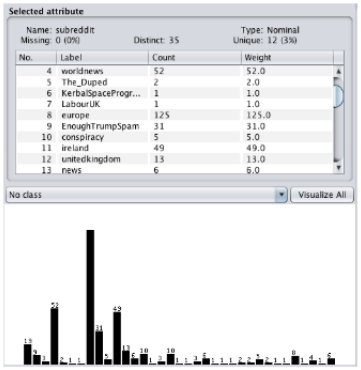

Figure 9: WEKA analysis of author 'AtomicKoala', including postings in other subreddits, and most frequently used words

Consider the analysis of the author 'AtomicKoala' in figure 8. This author has also contributed to the subreddits 'LabourUK','ireland', 'unitedkingdom', and 'europe'. Upon searching some of the user's posts, it was found that they refer to 'us Europeans', and 'here in Ireland'. This gives us an interesting insight into who might be using the r/politcs subreddit. Although the user posted mostly about the American election, 'europe' showed up in their most frequently

used words. An interesting thing about this user in particular is their interest in political parties. This user's frequent words included 'conservative', 'conservatives', 'democrat','democrats', 'left', 'gop' and 'party'. While we cannot conclude anything about European or Irish politics from this, it does show that the user talks a lot about the different parties, and that the party system is probably very important to how they view politics.

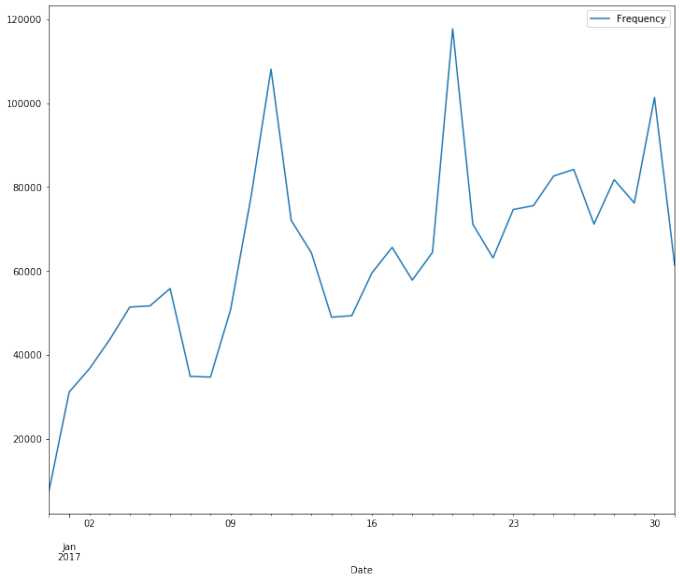

Figure 10: Number of Comments per day in Jan 2017 r/politics

6.6 Finding High Volume Days

A visualization of the time series dataframe shows the days of the month where users were most active in the r/politics subreddit. The mean value of comments posted per day were 63,000, with standard deviation 22,853 and a maximum value of 117,701. There are two peaks of very high activity (over 108,000 comments, or two standard deviations over the mean), namely on January 11th and January 20th. A simple search on major political events that happened on these days reveal that high user activity was correlated to significant political events. On January 11th, President Obama gave his farewell address, and Buzzfeed released a report with detailed unsavory information about Donald Trump. January 20th was President Trump's inauguration day.

Visualization of the time series of frequent comments (Figures 10/11) clustered on top users shows that posting trends of individual users have little correlation to the posting trends of the entire subreddit. Initially, an analysis of the most frequent users was conducted on a sample set consisting of the first partition of data (approximately 160,000 r/politics comments). Considering the new information from the time series analysis of comment density throughout the month, it can be concluded that the most frequent posters from a sample set is not representative of the aggregate set of most frequent users.

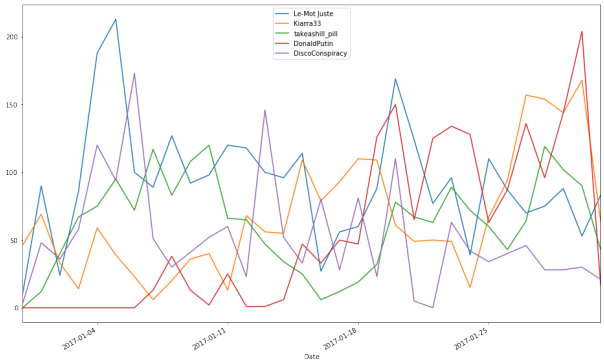

Figure 11: Top 5 Influencer Activity (1)

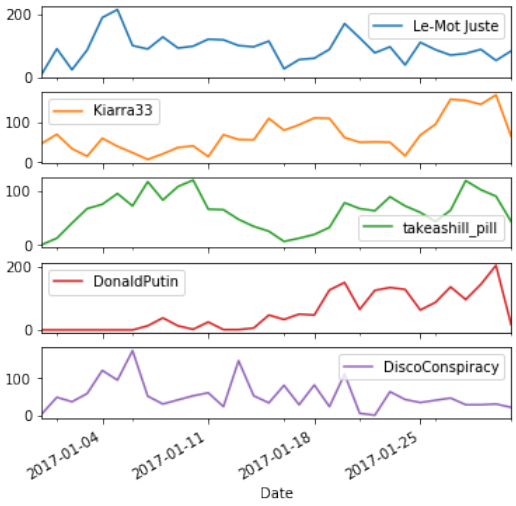

Figure 12: Top 5 Influencer Activity (2)

A comparison of the most frequent authors from different samples of the data set reinforce this conclusion. While certain frequent authors reoccur across the samples, it's evident that correlation of the sample data to the full set of data is low. Figures 12-14 demonstrate the differences in frequent author postings and statistics.

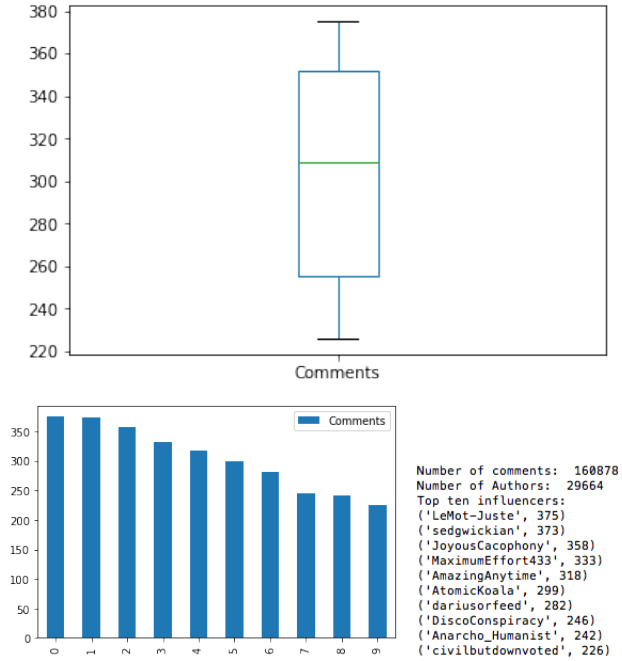

Figure 13: The top 10 influencers from the first partition of the data, with listed names, displayed on a bar chart. Statistical analysis including mean, standard deviation, and minimum and maximum values is also included

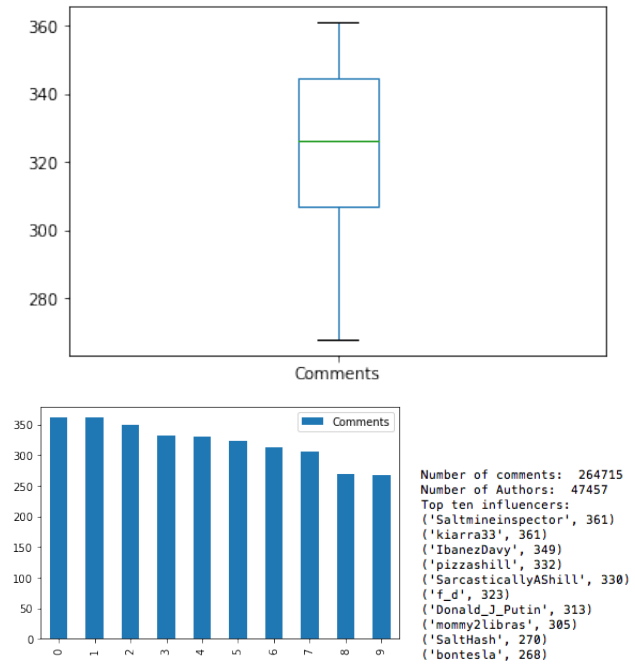

Figure 14: The top 10 influencers from the last partition of the data

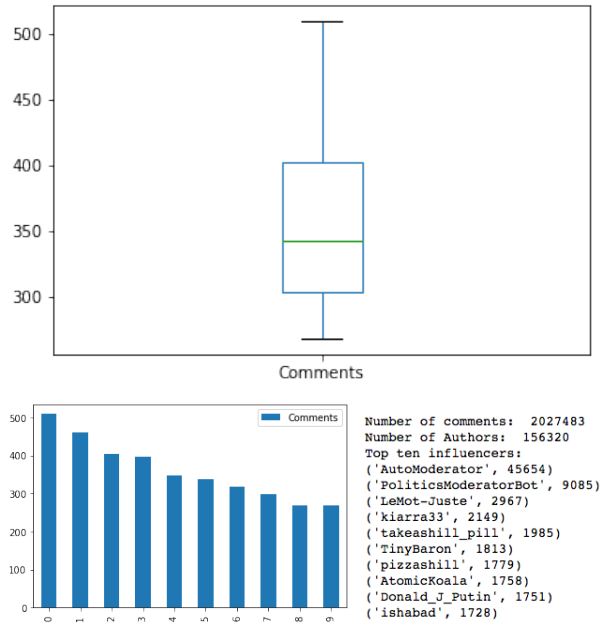

Figure 15: The top 10 influencers from the last partition of the data

6.7 Analysis of User Comments on Top Frequency Days

7. APPLICATIONS

This project could be used to determine which issues are most important to the users of Reddit. The website, largely considered an alternative new source, is notorious for the unregulated nature of its content. This type of data is valuable for analysis, as it could be seen as a more transparent connection to social trend, or the uncensored voice of the population that shuns mainstream news sources. While it is not practical to do an in depth analysis on every user on Reddit, even doing it on just a few shows some of what these users care about. Just as in the polls now, however, looking at only Reddit users may skew the data, and we have not done an in depth look on whether the users of Reddit are representative of a voting population.

Even so, knowing what the voters care about even in a smaller population could be valuable to politicians. Understanding the online community will become increasingly important in future elections and being able to speak to a majority of online users is already vastly important in many elections. Perhaps politicians will be able to use means like this to determine what policy they should be focusing on. While the r/politics subreddit is not a good place for local governments, analysis like this on subreddits such as r/Colorado might be a good place to understand user opinions.

Contrary to providing an unregulated data source for analysts and the like, identifying the most established users in a particular subreddit and analyzing their behavior could be useful for third parties seeking promotion in an environment free from advertisements. Established users with high karma and frequent activity have a lot of visibility on the platform. If said third parties could identify these users, they could offer incentive to these in exchange for a promotion of the

party's agenda. Using our chosen r/politics as a example, Clinton's political strategists could have promoted our Bernie Sanders supporter, kiara33, to post some pro-Clinton content in the r/politics subreddit, in hopes that this user's established presence would give this content high visibility.

Using subreddit posting activity, the significance of specific events to the Reddit community can be determined. A further analysis to text mine the comments posted on these days could provide insight into collective user sentiment regarding these events as well.

8. FUTURE WORK

This project lays a basis for a wide range of future work. More analysis could be done on user thoughts on certain topics. The work here looks at the user's word usage only, in future analysis, a user's feelings towards the topic could be studied. A user could mention a candidate a lot because they agree with that candidate's stance on a subject, or because they disagree strongly with them. The analysis we did will not show what a user thinks about a topic, only that the topic is important to them.

We could also find what the general populace of the subreddit thinks about a topic using the score attribute. Looking at the top used words in a high scoring topic could give more insight into how users react and think about a topic. With this kind of analysis, we do have to be careful that we are actually determining user sentiment and not that most users consider the comment to be a good point in a conversation.

Another interesting subject would be to determine whether smaller subreddits focused on a single candidate is comparable to the larger r/politics subreddit. Perhaps users who support one candidate are much more interested in different topics, and so we could build a network of different important topics. Using more of our time series analysis, we could determine what time of day the most posts occur. This could give us insights into who is posting in politics, we guessed mostly Americans, but one of the top posters, was from Europe, so that assumption could be completely inaccurate. This would also serve to give a rough estimate of where people have similar interests.

Doing a more in depth look at what the users care about could be done. This could involve looking at all the other subreddits the users who post in r/politics also post in. This would result in a network of subreddits.

9. CONCLUSIONS

In this paper, we show the results of our analysis of the r/politics subreddit. We wanted to address a broad range of topics, so several techniques were utilized to begin exploring this massive data set. To work with non-text attributes of this data, we first worked to reduce the large amount of data to a more manageable file size by eliminating all but comments from the r/politics subreddit. Reducing this data and preparing it for analysis made up a significant amount of work for this project, as processing times and software capabilities experience complications with large amounts of data.

Once reduced, this data was explored using various python libraries, along with further reduction techniques, data transformations, frequent pattern matching, and correlation analyses in an attempt to gain some knowledge from a large amount of information. From this, the most frequent users in r/politics were determined based on comment frequency, along with a time series graph which showed user activity per day. A closer look at the top users' posting trends showed no correlation between their most frequent activity to the aggregate activity of the subreddit. Counting the most frequent users in a sample wouldn't necessarily be representative of the population of r/ politics subreddits, at least for samples partitioned in evenly sized linear bins. As a whole, Reddit users in the r/ politics subreddit were the most active on days where significant political events took place.

The most common used words were unsurprising, including "trump and "president.

While looking at the top users and the comments they made, a better pattern emerged. When looking at a specific user, it became obvious what topics they cared deeply about. The top users had a wide variety of topics they cared about, but each user's top words seemed to follow a pattern, or contain more words from certain issues than other issues. These issues became even more clear when paired with the other subreddits a user posted in. There seemed to be no correlation between top users and score, however. Most of the top users had one comment that scored very high, and then most of the rest of the scores clustered around 0 or a low positive number. Most of the average scores were also between 10 and 0, showing that there didn't seem to be much of a difference between different user's top scores.

Data mining the r/ politics subreddit from January 2017 provided interesting and useful insight into its users' activity. It also provided us with a valuable experience of the data mining process, the tools and techniques involved, and the possible applications this process can provide.

10. REFERENCES

[1] Reddit january 2017 comments now available via torrent r/datasets, https://www.reddit.com/r/datasets/comments/5uftxe/reddit_january

[2] A. Barbaresi. Collection, Description, and Visualization of the German Reddit Corpus. In G. S. for Computational Linguistics & Language Technology, editor, *2nd Workshop on Natural Language Processing for Computer-Mediated Communication*, Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media, pages 7–11, Essen, Germany, Sept. 2015.

[3] M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity, http://ai2-s2-pdfs.s3.amazonaws.com/2db7/15a479c8961d3020fe906f7bedfa0311b93 2014.

[4] E. Gao. Analyzing trump v. clinton text data at reddit, Jan 2017.

[5] T. Weninger. An exploration of submissions and discussions in social news: mining collective intelligence of reddit, Feb 2014.