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. 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.

This project will revolve around Reddit, its information, and people that use it. We will be using this new found knowledge and the culture, the style, and the judgement of its users to lay the ground work that will support us and allow us to build the walls and the many floors of this project.

2. PROPOSED WORK AND CURRENT PROGRESS

2.1 Motivation

The aim of this project is to find find meaningful patterns in Reddit user comment data, and apply it to to solving one or more of several proposed problems:
-Determining which words and phrases were used most in political subreddits
-Cross checking users posting in the political subreddit with other comments and frequently used subreddits
-Identifying which cities are most frequently mentioned by users posting in the political subreddit
-Identifying user sentiment on political topics
-Determining whether there is any correlation between users, what places they mentioned, and how those places voted

These steps will help determine whether we can identify political leanings using Reddit. This is a relevant question because the recent election polls did not accurately predict results. While we are not planning to look at multiple months, this method could be used in the future to discern whether this 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.

2.2 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.

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, this information was applied 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 provide better assistance to these users.

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]. This paper 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.

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 a 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.

In ”Analyzing Trump v. Clinton text data at Reddit”, Emily Gao analyzed the comments from the r/politics subreddit 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 (words or phrases) 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.

2.3 Proposed Work

Several steps will have to be taken to pre-process and analyze this data. Integration with another data set is not necessary, as the data is contained in one large format JSON file. Although, in the future, we may integrate this data set with other sets with information on election results, location data, and the like. Upon review of the JSON formatted data, it is adequately cleaned, and none of the fields needs to be deleted or normalized. The dataset needs to be significantly reduced. For this project, only comments from specific sub-reddits will be used, namely, r/politics and related sub-reddits. Once reduced, this data will be moved to a relational database system for easier analysis.

After pre-processing, the comments' field 'body' will be mined to build a linguistic 'corpus', to identify significant words and phrases. This will include searching for names of people and places, patterns, and emoticons in the body of the comments. Data will be searched for users that post most frequently in the r/politics sub-reddit. After useful information is extracted from the data, a visualization of the geographic distribution of words and phrases based on commonly mentioned cities will be constructed.

Previous work on this topic used text mining techniques to analyze the frequency and use of the German language in various sub-reddits. The analysis of the r/politics sub-reddit was restricted to the month of July 2016, and focused on finding a correlation between the timestamp of a post and current events in the news. In this project, the focus will be on aggregate frequency of words and phrases in the r/politics and related sub-reddits using techniques employed to build the German linguistic corpus, rather than the correlation of word frequency to current events. This analysis could then be compared to previous work on the r/politics sub-reddit to compare trending topics before and after the election and concurrent inauguration.

Upon first attempts at pre-processing this data set, old questions can be addressed, along with new questions that were posed after initial analyses. One possibility is the development of a program with python tools capable of processing the large body of comments from the r/politics subreddit (over 1.5 GB). Another possibility is working with the current program, which functions on data set under 40,000,000 comments, and simply partition to data into two sections and average the results of these partitions. Work on weka now includes building the corpus successfully, and extracting information and knowledge from this corpus once this is resolved.

Unanticipated complications needs to be addressed; upon reducing the data to the r/politics subreddit, and extracting the most frequent comment authors, it became apparent that there are a large number of fake accounts (”shill” accounts) that skew results. This is a significant discovery and should be addressed accordingly, so proposed work now includes coming up with a method to deal with these accounts.

2.4 Data Set

The data set being used was found on the reddit website[1]. It is approximately 42GB of user comments from the month of January 2017. The data is stored in JSON string format; each object represents a comment. There are 78,946,585 comments in the data set and each object has the following potentially useful fields: Body, edited, Author, Subreddit, Subreddit ID, Author Flair, Stickied, ID, parent ID, score, retrieved on, Controversiality, link ID. The body contains the actual text of the comments, which will be used to mine for frequent words and phrases. The Subreddit and Subreddit ID will be useful for our initial data reduction. The Author field stores the username of the author of each comment. This could be used for tracking users to determine the types of users likely to post. We may look at the score, which tracks the upvotes and downvotes, to determine how popular or unpopular an opinion is.

2.5 Evaluation Methods

We will not have to clean our data, since all the fields are filled with data that will work for us. Although initially we did not think we needed to remove any comments, we have since discovered our data contains many redundant comments. These come from bots, deleted or removed comments. Since these comments do not add any information on user ideas, we will remove these comments so we are only looking at data that matters to our questions. We will need to reduce our data into a more manageable size. The first way is to make sure we are only looking at the comments written in English. Once we have those comments we will split on the sub-reddit or sub-reddit ID field, looking specifically at r/politics and related sub-reddits. Once we have this smaller data set, we will use sampling to determine common words and phrases we will be looking for. Although sampling may not catch all of the frequent words and phrases, we gain a good understanding of what we will be looking for in our complete data set. Once we go through and determine which tokens we are looking for, we can remove any comments that do not contain any of them. In this way, we can again reduce the number of items we are looking at. Once we have our reduced data set, we can find the frequency of any tokens found as well as determining the sentiment of the user's comment.

After this, we will again look at our whole data set to determine what other sub-reddits the users were posting, or subscribed to. This will give us more information about the users themselves. From here, we will look at the posts of these specific users and determine what places they mention. We may try to remove any data mentioning ”travel” or places mentioned in traveling sub-reddits. We will need to do some normalization here since cities are more likely to be mentioned than towns, and cities could have different political views from the surrounding towns. We will integrate another data set with election data at this point and try to find correlations with user sentiment and how the area voted.

Evaluation methods already used include string and pattern matching in weka to build a corpus of the most influential words in comment bodies. A similar pattern matching was also used in python to reduce the data to the correct subreddit, and then to count the the frequency of comment authors in this subreddit. Python's collections library was used to determine the "Top Ten Influencers" in r/politics, by employing an (O(nlogn)) sorting algorithm and a minimum threshold of influence.

Future evaluation methods will include advanced clustering techniques to determine correlations in the text corpus that could lead to meaningful knowledge extraction. Clustering techniques may also be employed to determine the likelihood that a comment comes from a "shill" account, and potentially to build a network of author interaction by tracing comment/reply id's (the parent id of each comment is included as a field in the JSON data).

2.6 Tools

2.6.1 JSON

Our dataset is JSON formatted. The JSON file data is simple and readable, comparable to a python dictionary.

2.6.2 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. Currently, python's json library has been sufficient for this project. The Pandas Library contains various Python specific tools for data analysis and data display with the convenience of easy cooperation between the Python scripts. Depending on the ease of working with the data as the project progresses, a switch to Pandas may be utilized to improve efficiency and provide data visualizations as well.

2.6.3 WEKA

WEKA was used to count frequent words in the comment body field. It allows for easy text collection and organization text. 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.

2.6.4 Database

Another possible tool would be to insert the data set into a database or data warehouse such as MySQL or BigQuery. This would give us an easy, simple way to query our data for specific information. This option will be explored if more complications arise using the previous tools mentioned

2.7 Milestones

2.7.1 First Milestones

1. Data Integration and pre-processing
   reduce the JSON data to include only comments from the r/Politics sub-reddit. Transform the data to a mySQL or BigQuery database for analysis.

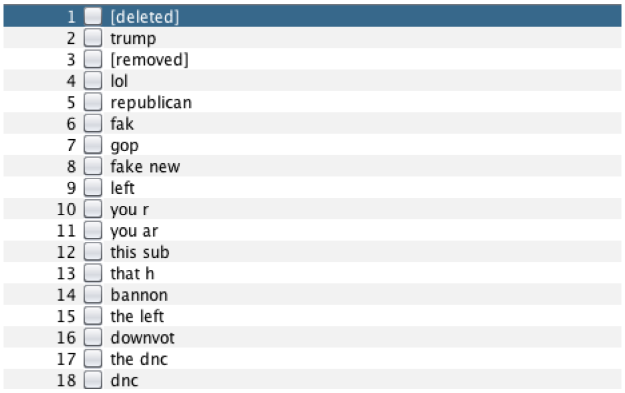
2. Process for the Data
   Information extraction of the new dataset. Create a corpus that contains the most common words and phrases used in comment bodies. Identify the most active users in the r/Politics sub-reddit based on the number of comments and posts they aggregate. Identify the most popular comments based on the number of upvotes received. Search for frequently mentioned cities, and words associated with these cities.

3. Evaluation
   Search for a correlation between most frequently used words and phrases and events in the month of January 2017. Use this and most popular comments to identify the most popular topics on the sub-reddit this months, and predict a collective user opinion about them. Map a geographic distribution of reddit comments based on content, by cross comparing with geographic location of mentioned cities, to identify a correlation between comment topic and location or demographic in the United States. Identify significant interactions between sub-reddits and additional user characteristics, by tracking the activity of the most active users in the r/Politics sub-reddit to other sub-reddits. Potential visualization of data to help better understand results.

2.7.2 Milestones Achieved

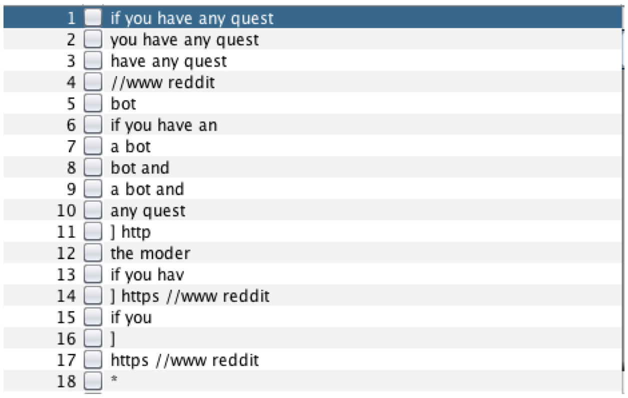
1. Data Pre-processing and Cleaning
   Using a simple Python script and the JSON library, a script was developed to reduce the JSON file to a more manageable size. This script searches the entire 42GB JSON file for a certain cindition (In this case subreddit == politics), and transfers the selected comment nodes to a much smaller JSON file. The script was successfully tested on a sample size of 100,000, but the json.dump() method used to create the new file experiences complications for dumps above 1.5 GB. After more testing, a JSON file was created successfully by partitioning the original data file into two sets of 40,000,000 comments each. Additional analysis can be performed by two team members concurrently (this will significantly reduce time spent) and results can be aggregated after.

2. Data Processing
   We utilized WEKA to text analyze the dataset and extract useful information. Figure 1 shows how we were able to extract the top words used in the r\Politics subreddit comments. The results were as expected considering the events that occurred in January 2017. Political words relating to the US presidency were among the top. Words such as "trump" and variations of "fake news" were all among the top words. Additionally, Reddit keywords "[deleted]" and "[removed]", which were irrelevant, were also among the frequent words used. This shows we will have to come up with a better method to deal with these comments. Next we used WEKA to analyze the comments and score relationship. Image 2 displays the ngrams that are most



1	□	[deleted]
2	□	trump
3	□	[removed]
4	□	lol
5	□	republican
6	□	fak
7	□	gop
8	□	fake new
9	□	left
10	□	you r
11	□	you ar
12	□	this sub
13	□	that h
14	□	bannon
15	□	the left
16	□	downvot
17	□	the dnc
18	□	dnc

Figure 1: Most Useful Words to Split on When Determining Difference Between Negative Comment Scores



1	□	if you have any quest
2	□	you have any quest
3	□	have any quest
4	□	//www reddit
5	□	bot
6	□	if you have an
7	□	a bot
8	□	bot and
9	□	a bot and
10	□	any quest
11	□	] http
12	□	the moder
13	□	if you hav
14	□	] https //www reddit
15	□	if you
16	□	]
17	□	https //www reddit
18	□	*

Figure 2: Most Useful NGrams to Determine Score

useful in determining whether a comment had a score less than -20. Most comments in the negatives would score between -1 and -20. So we used this knowledge to show the comments with a lower amount of votes and compared them to each other. Using simple Python script, we were able to search the top authors in the r/Politics sub-reddit. Our results had many errors due to Reddit bots.

3. Evaluation
   We searched for a correlation between most frequently used words and phrases and events in the month of January 2017. Using this and the most popular comments to identify the most popular topics on the subreddit this months, and predict a collective user opinion about them. Map a geographic distribution of reddit comments based on content, by cross comparing with geographic location of mentioned cities, to identify a correlation between comment topic and location or demographic in the United States. Identify significant interactions between sub-reddits and additional user characteristics, by tracking the activity of the most active users in the r/Politics sub-reddit to other sub-reddits. Potential visualization of data to help better understand results.

2.7.3 Future Plans

1. Data Processing
   Discovery of a large number of ”shill” account comments gives rise to a new host of questions. There are enough to significantly skew the results of analysis; this is a serious consideration when the purpose of analysis is to determine user behavior, as the addition of automated and false accounts will make any results inaccurate. Thus, an appropriate method to handle these shill comments is needed for an accurate extraction of knowledge. Should the false comment simply be removed? This could be achieved with by setting certain flags in a search as a minimum threshold for detection of a false account. It would also be valuable to compare an analysis including the fake accounts with an analysis after they are removed to determine the influence that shills have on the data.
   Figure 2 above is just an example of using weka to remove shill comments. This could also be achieved in python.

2. Data Storage and Display
   So far, an optimum storage toll hasn't been chosen. The Pandas library is in consideration, as are several relational database services that could make querying data easier, namely mySQL and SQLite. At this point, data is simply being moved to a smaller JSON format file.

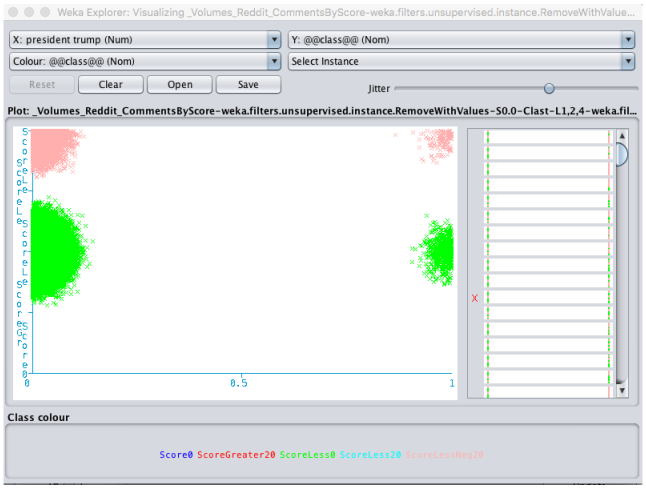
3. Further information Automated comments and ”shill” accounts
   Upon further research, it was discovered that shill accounts are a large problem in the reddit community, and the only method available to detect them are volunteer reddit users that search for them manually. One member of our group reached out to the r\Politics admin named in a forbes article concerning the issue. An interview with this admin is scheduled; he has agreed to help determine the parameters that would make an automated system to detect shill accounts more efficient.

3. CONCLUSIONS

Using WEKA we managed to see some interesting relationships in the r\Politics subreddit. Firstly, we looked at commonly used stem words and compared the score when they were used. Figure 3 shows the relationship of the stem word “trump” and Reddit score. Because most comments reside between a score of zero to ten, we decided to only look at those comments scored below zero and above negative twenty. The data shows that a negative score is more prevalent when “trump” or variations of that word is used in a comment.

WEKA also allowed us to compare correlation between the score and when two different ngrams were used in the same comment. In figure 4, we looked at comments when both “trump” and “obama” were used. The graph shows that when both ”trump” and ”obama” were used, the score was often only in the range from -1 to -20. However, when only ”trump” was used the comment was much more likely to be less than -20, and when only ”obama” was used the comment was most likely in the -1 to -20 range.

We can look at specific non political words as well. Figure 5 and figure 6 show the scores of comments based on usage.



Weka Explorer: Visualizing _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValue...

X: president trump (Num)

Y: @@class@@ (Nom)

Colour: @@class@@ (Nom)

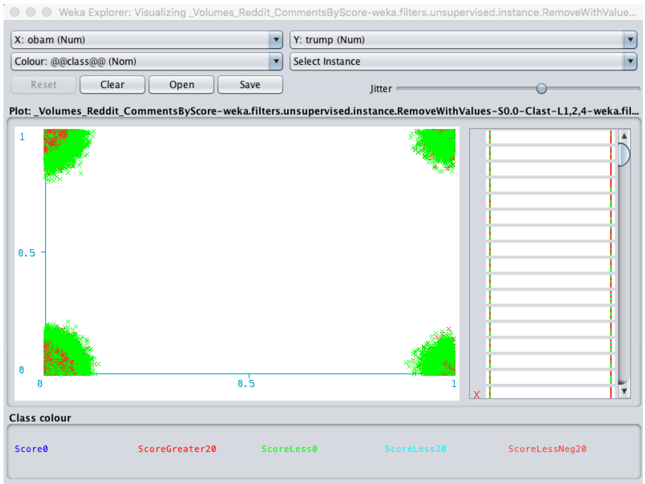
Select Instance

Plot: _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-L1,2,4-weka.fil...

Class colour

Score0 ScoreGreater20 ScoreLess0 ScoreLess20 ScoreLessNeg20

Figure 3: Score When ”president trump” was used



Weka Explorer: Visualizing _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValue...

X: obam (Num)

Y: trump (Num)

Colour: @@class@@ (Nom)

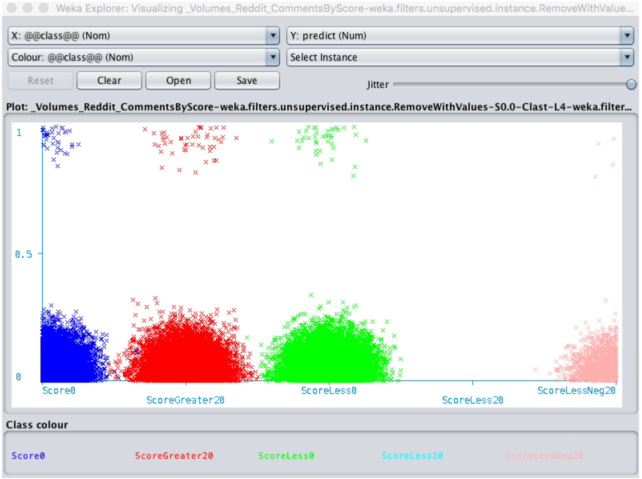
Select Instance

Plot: _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-L1,2,4-weka.fil...

Class colour

Score0 ScoreGreater20 ScoreLess0 ScoreLess20 ScoreLessNeg20

Figure 4: Score When ”trump” and ”obama” was used



Weka Explorer: Visualizing _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValue...

X: @@class@@ (Nom)

Y: predict (Num)

Colour: @@class@@ (Nom)

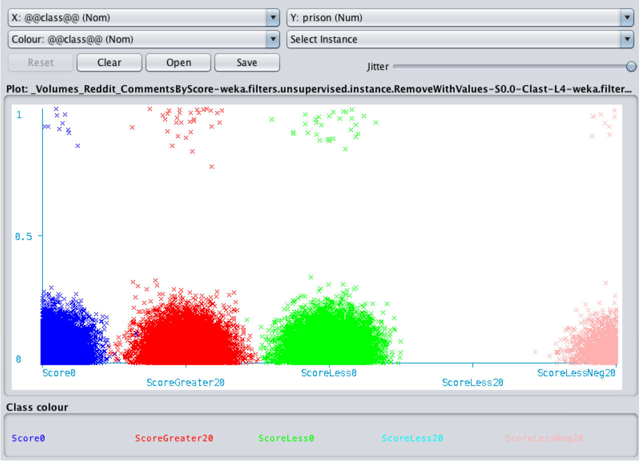
Select Instance

Plot: _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-L4-weka.filter...

Class colour

Score0 ScoreGreater20 ScoreLess0 ScoreLess20 ScoreLessNeg20

Figure 5: Score When ”predict” was used



X: @@class@@ (Nom)

Y: prison (Num)

Colour: @@class@@ (Nom)

Select Instance

Plot: _Volumes_Reddit_CommentsByScore-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-L4-weka.filter...

Class colour

Score0 ScoreGreater20 ScoreLess0 ScoreLess20 ScoreLessNeg20

Figure 6: Score When ”prison” was used

A comment shows that the word was used if it is marked at the vertical point one; a zero if it does not appear in the comment. As can be observed in the graphs, "predict" seems to be most active in those comments with a score greater than twenty, while "prison" is a little more difficult to differentiate.

Through our mining of Reddit, we hope to gain a more complete view of political sentiment. This will be accomplished using data mining techniques and text analysis. Reddit data aligns perfectly with this as we can search through the political sub-reddits and have plenty of data to find meaningful results.

Analysis of the comment authors revealed unanticipated information about the data, and introduced new complications to the analysis of this data. Determining the top influencers and separating these from the false accounts could provide insight into the inner workings of reddit, and conclusions could show that these shill accounts have a greater influence on the site than previously suspected.

We plan to use these political sub-reddits to determine whether we can gain any useful knowledge about a user's political ideas and approximate location based on what they post on Reddit. Reddit contains a diverse group of users from different political groups. This results in a large dataset that could be used later with future comment data from future elections and political events to determine whether we could better predict the outcome using this analysis.

4. REFERENCES

[1] Reddit january 2017 comments now available via torrent r/datasets, https://www.reddit.com/r/datasets/comments/5uftxe/reddit_january_2017_comments_now_available_via/.

[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/2db7/15a479c8961d3020fe906f7bedfa0311b937.pdf, 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.