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 it’s 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.

2. BODY

2.1 Motivation

The aim of this project is to find find meaningful patterns in Reddit User comment data, and apply 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.

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 is 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 ’disinhibiton’ in the Reddit community due to analyze its effectiveness in getting users to disclose thier 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. These 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 subreddits[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 subreddits, 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 dataset for the most frequently used tokens (words or phrases) used in the subreddit 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 preprocess 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 subreddits will be used, namely, r/politics and related subreddits. Once reduced, this data will be moved to a relational database system for easier analysis.

After preprocessing, 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 subreddit. 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 subreddits. The analysis of the r/politics subreddit was restricted to the month of July 2017, 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 subreddits using techniques employed to build the German linguistic corpus, rather thatn the correlation of word frequency to current events. This analysis could then be compared to previous work on the r/politics subreddit to compare trending topics before and after the election and concurrent inauguration.

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. Each file from the JSON is a single comment, so we do not have to deal with redundant data of any kind. 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 subreddit or subreddit ID field, looking specifically at r/politics and related subreddits. 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 subreddits the users were posting. 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 subreddits. 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.

2.6 Tools

Our dataset is JSON formatted, which will work well with Python.

To do our language detection, we will use a tool such as enchant or langid.py

Our data will be stored in either BigQuery or MySQL

We may use WEKA for looking through our data

2.7 Milestones

1. Data Integration and Preprocessing
   reduce the JSON data to include only comments from the r/Politics subreddit. 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 subreddit 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 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 subreddits and additional user characteristics, by tracking the activity of the most active users in the r/Politics subreddit to other subreddits. Potential visualization of data to help better understand results.

2.8 Summary of Peer Review Session

The peer review session helped us narrow down what tools we were going to use as well as helping us get a clearer idea of what our proposed work would be. Before the peer review session, we were planning to work solely from our raw data with python. But after seeing many of the teams mention using a databases, we realized this would be a much better way to store and access our data. We also gained a better understanding of what types of questions we could be asking with our data. Seeing the types of questions other teams were asking helped us narrow down what we were asking. Teams that helped with this were specifically the teams doing other text mining, such as the teams mining the yelp reviews. This was both a good thing and a bad thing since once we saw others proposals, we had so many ideas it was hard to narrow down!

3. CONCLUSIONS

Through our mining of Reddit, we hope to gain a more complete view of political sentiment. This will be accomplished using datamining techniques and text analysis. Reddit data aligns perfectly with this as we can search through the political subreddits and have plenty of data to find meaningful results. We plan to use these political subreddits 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. This research could be used later with data from different elections and political events to determine whether we could better predict the outcome using this data 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.