Statistical Analysis of Comments on Reddit

Saurabh Puri
srp3858@rit.edu

Prof. Carol J. Romanowski
Rochester Institute of Technology

Abstract:
Social networking site is a type of repository where you can find all type of information from all around the world. Reddit is a rising social website containing meaningful and useful information. Various view, comments and likes by people on social networking sites enables them to gather a large set of data to arrive at some consensus or analysis. This paper is a study to analyze the tendency of human to stick to a topic while commenting on a social networking site.

I. Introduction:
Reddit is an entertainment, social news networking news website where registered community members can post their content like post, links are categorized as per their interest area to discuss with each other. It has numerous popular website all throughout the world specially United States. People uses this medium more often to communicate and share news through their reddit blogs. One can register their account on reddit for free without the need of any email address. After logging in reddit, people can cast vote on the posts and comment to highlight the topic or link posted. Creation of one’s own subreddit on a topic is also possible by subscribing to its front page.

With a large number of users posting and commenting their opinions and posts in the form of subreddits, one can analyze various trends and correlation between them. It becomes an essential task to capture interesting patterns from the idea of the topic posted by the users about a relevant subject. The ability to comment on subreddits in order to voice one’s opinion and to share details about it makes it an interesting bet among the masses. Moreover, these comment on Reddit is mostly public allowing people to view the discussion without any restrictive access. Due to this open source kind of property, it is most widely accessed forum across the countries.

The main objective of this project is to determine how much time it takes to divert from a topic while commenting. Thus, the main task is to fetch a post along with all the comments of that post to determine the outcome. The outcome depends on the capability of people to stick to the same post derived from the original topic. People can use this platform to check the connectivity between the post updated and detailed information that could be obtained from comments posted for each subreddit. This could make it an intelligent forum to check the legitimacy and credibility of
opinions and posts shared over social networking websites which can further result in stronger information sharing among a large number of people connected globally.

This capstone project is documented in the following manner. Section 2 gives a description of the work done by others regarding this project and also provides a brief understanding of their work. Section 3 contains the description of the algorithms used during the entire phase of the project and step by step detailed implementation of the project. Section 4 contains the results obtained from the implementation and its visualizations and also discusses how the conclusions were drawn from the visualizations and which parameters were used to do so. Section 5 includes the future work which can be done to extend the project.

II. Previous Work:

A lot of work has been done focusing on the comments section of Reddit but it provides more detailed explanations about the users who are commenting. The relation between the post and comment is not targeted to study previously. Social networking sites connect the whole world together. Which further brings a lot of data in one place and it triggers researchers to analyze and study different patterns surrounding it. People can post and share news articles which are posted by someone else to increase the trendiness of that topic. Reddit is information rich massive repository of valuable knowledge on the diverse range of topics. Millions of web users leverage this platform to post, view and discuss almost anything.

In “Information and Social Analysis of Reddit”, Troy Steinbauer [1] provides in-depth analysis by providing the structure of the comments section of reddit. Through this structure, relationship between users can be seen. Which user communicates on which post and how often does he comment, these structures can be studied with the help of this paper. Reddit being a rising social news website, this paper provides deep analysis of Reddit dealing initially with data crawling technique along with describing the relation between the article and its sub-sections focusing primarily on comments.

In the paper by Tim Weninger [2], the area of study and research focuses on the hierarchical structure in the discussion threads. They focused on three main topics for research. First, one is to analyze how deep does the discussion on the threads resembles like a hierarchy. Next is to check whether through these comments they can enhance the searching of data on the web. The last area of concern is to find out the areas in which the highest scoring comments reside. It provides a detailed explanation of different parameters in Reddit like comments, post, voting, subreddits, and karma. This paper does not focus on crawling of data, instead provides structure wise explanation on the data from comments on reddit. It provides visualizations and description of the topical hierarchies of comments. Through mathematical proof, the author has provided an explanation as to how the general search can be improved through comments. The main aim was to derive the comment value which is the key to all the visualizations.

Salman Jamali (2009) & Rangwala [3] focus on individual user behavior. The participation of the user in the different topic is analyzed. Digg is another social networking website like reddit. The authors work predicts popularity for a particular content which is linked using Digg. Classification and regression framework was used to get these results. More focus was given in user characterization to know which sections does that user comments on.
Authors Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit [4] has provided with a step by step guide for using nltk in python language. The author gave a description of what is NLP and how it can be used to find the percentage of similarity between words. How text can be processed through nltk and which all different functions can be performed on the text is shown. The work also included Corpus module which is very widely used when large chunks of data are present in the dataset. Corpus uses raw data and many functions can be performed through it. Wordnet is another module in which a structured and semantically oriented English dictionary is present which includes synonyms, antonyms, hyponyms, depth of a synset and trees. Tokenization can be achieved easily through nltk but for that format of data should be known to the user. The author also provided details about Stemming and lemmatization. This work helped me research and study in depth of wordnet in nltk. The concepts are linked together in a hierarchy know as root synsets which make it easier to navigate between the different functionality. Through import statements, the functions can be imported. Focus was also put on synsets.

Lakkaraju [5] evaluated Reddit submissions based on the title of the submission. In this paper relationship between title, submission count and choice of content is described. Also, its impact on Reddit submission is analyzed. A model was created for defining a language which included good as well as bad words, performs speech tagging and performs sentiment analysis which helped to know the success for every submission which made the title more famous. The language model was not only used for knowing how successful the post is but other attributes like content quality, submission counts, and a number of comments associated with each post were also taken into consideration which contributed towards the credibility of the content on social news websites.

III. Implementation:

Above figure displays the flow diagram of the implementation. The whole implementation is divided into checkpoints. It starts with data collection which involves connecting to Reddit and fetching a post with all its comments. Next is Data cleaning and validation which involves clean raw data to usable form and analyzing the data. Similarity check consists of comparing the comments to the post and saving the results in a tree structure. Evaluation step includes processing the tree and displaying the final result.
Data Collection:

Reddit involves thousands of subreddits which makes it difficult to focus on a particular post while fetching it. I focused on popular subreddits to start with and then kept of increasing the scope as the checkpoints in the project were getting completed. Through research and analysis, came across points in which the author faced problems while trying to fetch the data from Reddit through the API. There were restrictions on the number of data that could be fetched in a particular period of time and if not adhered to, then the account was blacklisted. Taking that into consideration, restrictions were put for fetching in a particular time period. Researched about API's which connect to Reddit. There were JRAW, etc which are unstable. Stabilized on PRAW API as it is the only stable API for python and easy to implement. Researched about API's to connect to and found out PRAW API which has a detailed explanation on how to connect to Reddit. Installed praw using pip command of python package. Below is the screenshot of its installation.

![Screenshot of praw installation](image)

Praw has many user agents through which Reddit can be accessed. A user-agent is a piece of code which adheres to a particular set of rules of an application. After analyzing different user agents and their rules, finalized on "Karma breakdown 1.0 /u/_Daimon_" where Karma breakdown is the bot name, 1.0 is its version and /u/_Daimon_ is the username who created this user agent. The praw object can be used to get all the comments of a particular post in raw form, the creation time for the comments, etc. The title parameter of the object consists of the text data.

After importing praw package and login into my account of Reddit through praw, fetched data from a test subreddit initially to make sure if the program is working correctly and no rules are violated. Every individual post has a submission number to it which makes it unique. Through that submission number, the post can be accessed using Praw object. The comments are stored in JSON format. So after iterating through the JSON file using praw object, every individual comment is converted to a string so that it can be stored in a file. Challenges I came across during conversion of JSON object to a string and saving it in a file was that the conversion required asci code. The
rest of the format were ignored. Other formats include links and images. Another problem I faced was that the comments were saved only till a certain page limit but when load more comments were clicked then the program used to fetch more data. This problem was addressed to by adding a piece of code: "submission.replace_more_comments(limit=None, threshold=0)". submission is the object which consists of the reference for the particular post targeted. Limit = none ensures that all the comments are fetched within the time frame of 30 seconds. The id of all the comments is saved in the set. This parameter makes sure that the same comment is not saved twice in the file.

One of the most unavoidable situation was during connectivity of reddit server. When the server was down, more than one attempts was made for crawling data as web crawler was not able to distinguish between reddit being down and a bad url. Had put a condition which would check for login successfulness and proceed accordingly. If this was not handled then chunks of data could get missing between the crawling processes.

b. Data Cleaning and Preparation:

After saving all the comments in a file, analysis was performed on how different data comments were saved. Created a list of stop words which will be used to remove those words from the comments for better accuracy in terms of similarity. Tokenization was performed so that every word can be treated individually. Tokenization helps to validate the data properly. Ran all the words in a loop to remove special characters. But kept single apostrophe, so that the words make sense while checking for similarity. The comments were then read from the file and saved in a list for further comparing. The original post text was fetched using praw object. The submission number for the post was used to make a reference and then the title parameter was used to fetch the text of the post. This text was also saved in a list after being cleaned and divided into tokens.

c. Similarity Checking:

Topic modeling was an option which could be used for similarity checking but through research came across natural language processing toolkit(nltk). Nltk is a collection of libraries and programs for natural language processing for python programming language. Wordnet is a part of nltk which helps to check the similarity between words. Wordnet provided better results. Had created a sample data whose output was known to me. Tested the program for this data initially. Topic modeling didn't provide better results than wordnet of nltk. Implemented different ways for checking the similarity but the best accuracy was achieved from wordnet of natural language toolkit package. Wordnet is used to find out the sense of similarity between two words. The sense of both the words are compared and the result is given in the form of number ranging from 0 – 1, where 0 is the least possible match and 1 being the highest. For all the comments, timestamp was fetched. The challenge in fetching the time was that the format was in Unix time. It took me a lot of time to analyze the time format as no documentation was provided for it.
d. **Outcome:**

![Ideal Tree](image1)

![Drifting Tree](image2)

The results of all the comments were stored in a tree. Then the whole tree was compared to the ideal tree. How the tree should look when people are not drifting away is the representation of the ideal tree. It provides a basis for comparing the tree resulted from the comments. Above figure gives a visual representation of how the tree looks like. The drifting tree is the output individual comments compared to its post.

![Sub-Reddits](image3)

Tested the accuracy for over 50 posts manually and then finalized on 0.43% as the distinguisher between drifting and not drifting. After accuracy was fixed, ran the code on hundreds of posts. In the end, did the analysis for all the results of the data and applied k-means clustering on it. Above
figure gives an average number of comments fetched in a particular section of subreddit. These subreddits were the ones which I targeted the most and results were derived from them.

IV. Results:

<table>
<thead>
<tr>
<th>Sub-Reddits Targeted</th>
<th>Avg. no. of comments</th>
<th>Drifting/Not-Drifting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jokes</td>
<td>500</td>
<td>Not-Drifting</td>
</tr>
<tr>
<td>News</td>
<td>1500</td>
<td>Drifting</td>
</tr>
<tr>
<td>Programming</td>
<td>200</td>
<td>Not-Drifting</td>
</tr>
<tr>
<td>askReddit</td>
<td>2500</td>
<td>Not-Drifting</td>
</tr>
</tbody>
</table>

Most of the post did not drift. The ones which did took a lot of time to deviate. This was not what I expected and the results were a surprise to me. Through this project, the relevance between Reddit post and its comments was found its content can be used as a real-time performance indicator. By evaluating the relevance between the original topic and the subject area discussed in comments, the credibility of the social news networking website was projected.

V. Future Scope:

The project is built to analyze textual comments in relevance to the original post and find the relation between them. It cannot handle direct links in the comments to analyze the content posted in that link. Due to heterogeneity and lack of proper structure in direct links, new specialized mining algorithms need to be written supporting web mining. The tree structure can be analyzed in-depth for its behavioral meaning. It was out of the project scope to read the content available in the link for checking its relevance with the original post. Different comparing techniques can be used to check for better accuracy.

According to socioeconomics, social mood is a key driver of political, social and economic events. This project though did not dive into accessing the social mood or performing sentiment analysis of people on Reddit. It is difficult to classify textual comments by genre, style, author or sentiment orientation because the classification has to be done regardless of the topic of the document which widens the scope of the project. This implementation can also be done for Facebook and results can be derived for the similar topics targeted in Reddit. Thus, a comparison can be done to show is it the same for all the social networking sites or it is just for Reddit.

A crisis by nature is uncertain, often featuring a high sense of urgency which can feature either delayed or inaccurate coverage through social media news. As a consequence, it becomes exceptionally difficult to understand the potential impact on events. Therefore, a better
understanding of the topic through this forum would be further beneficial to create situational awareness for an effective crisis management by assessing and ranking information in terms of its quality, credibility, and relevance.

VI. References:


