MINING DATA FROM REDDIT

GROUP D

Abdul Gafar Manuel Meque

Chih-Ming Chen

Sachit Mahajan

Introduction

The purpose of this report is to show how to mine data from Reddit which is our main data source.

Goal

The goal of this project is to find the relationship among scores (higher/lower), post emotion/polarity under different subreddits. Also focus is to find the words that can affect the score in certain subreddit.

Methodology

In this project, according to our intended goals, we broke-down the entire process into **N** steps. Each of the steps is designed to address a smaller subset of our problem, and they are depicted in the work-flow diagram below.

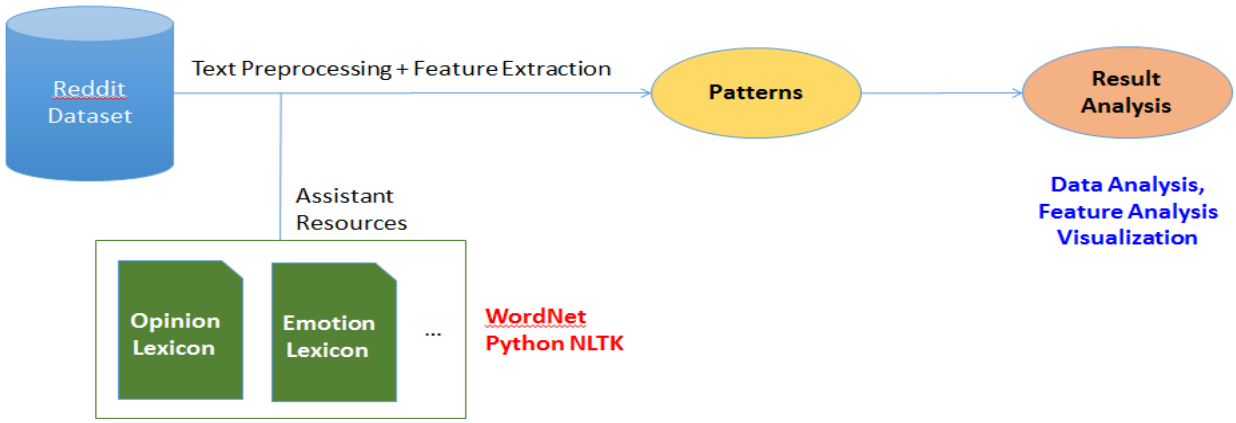

Figure 1: Work Flow

Application Domain and Data Familiarization

In the beginning of the project, reddit platform was somehow unknown to the group members, so was the data. For the first week each group member engaged in get acquainted with various aspects of reddit, all its features, rules and regulations, so we could better understand the data at hand.

The dataset is comprised of 54504410 comments, a fraction of all comments made on reddit during the month of May 2015 available publicly, organized in a single table 'May2015' and provided by (Kaggle, 2015) .

Each row in the table May2015 represents a single comment, and has the following attributes: ("gilded","author_flair_text","author_flair_css_class","retrieved_on","ups","subreddit_id","edited","controversiality","parent_id","subreddit","body,"created_utc","downs","score","author","archived","distinguished","id","score_hidden","name","link_id"). Basically for each comment we have details such as the

content, the comment author, subreddit, date created, date retrieved, score (the difference between upvotes and downvotes), the number of upvotes, and some more, not relevant to the current work.

Preprocessing

Step 1

After having a better understanding of the dataset and the domain, and based on our project goals discarded the attributes that were deemed not relevant, and settled on using only body, subbredit and score.

Step 2

Since we work on the sentiment/emotions behind the reddit text, we used language as a parameter for further reduction of the dataset. Using TextBlob a tool from (Keen, et al.), we removed all non-English comments.

Step 3

As another dimension reduction parameter, we define a threshold of a minimum of 50000 comments for a subreddit to be considered popular, so we removed all comments from unpopular subreddits.

Step 4

Using TextBlob by (Keen, et al.) we classified all comments as being either of positive, negative or neutral sentiment.

Step 5

In the 5th step we applied a text processing tool in TextBlob to retrieve noun-phrases, sentiment lexicon in each of the subreddit.

After all five steps where concluded our final dataset had the following structure:

Subreddit name	Score	Noun phrases	Negative words	positive words	Sentiment label

FEATURE EXTRACTION

Step 1

To ensure the quality of extracted words, we apply term frequency to each commends as a kind of confident level filtering mechanism. In this report, we set the threshold with 32, which means we remove the words appeared in less than 32 distinct documents.

Step 2

In descriptive statistics, we calculate the first quartile and the third quartile to realize the distribution. Note that all the calculated scores are scale by the logarithm. This way avoids the outlier score issue. In addition, we found most of comments fall into zero score so that we focus on investigating the words lead to the highest/lowest score based on only positive/negative score comments.

Step 3

We finally sort the words by the first quartile and the third quartile to get the influential words, and plot them in the box-plot figures.

Results

Some of the results obtained during the processing part have been shown below.

Figure2 and Figure 3 show the positive and negative words distribution which is obtained during the feature extraction step. Figure 4 illustrates the overall distribution for the detected sentiments.

Figure 2: Positive Words Distribution

Figure 3: Negative Words Distribution

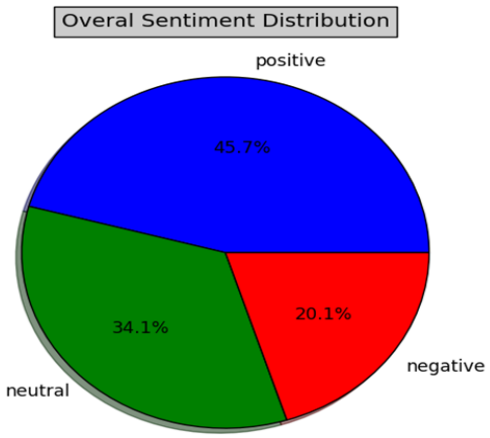

Figure 4: Overall Sentiment Distribution

Final Outcome

We present and analyze the results from 5 different perspectives with corresponding box-plot figures. The X-axis is the exponential score value. The Y-axis represents the exact words. The subreddit name is listed in top right corner.

1) We first show that the overall impact of a word could be positive or negative, as the shown in Figure 5 and Figure 6. In average, there are more than 99% of words are in the zero mean distribution. However, it is clear to see that all of the extracted words in Figure 5 and Figure 6 can receive a score away from the zero. This result implies that the words indeed bias the scores.

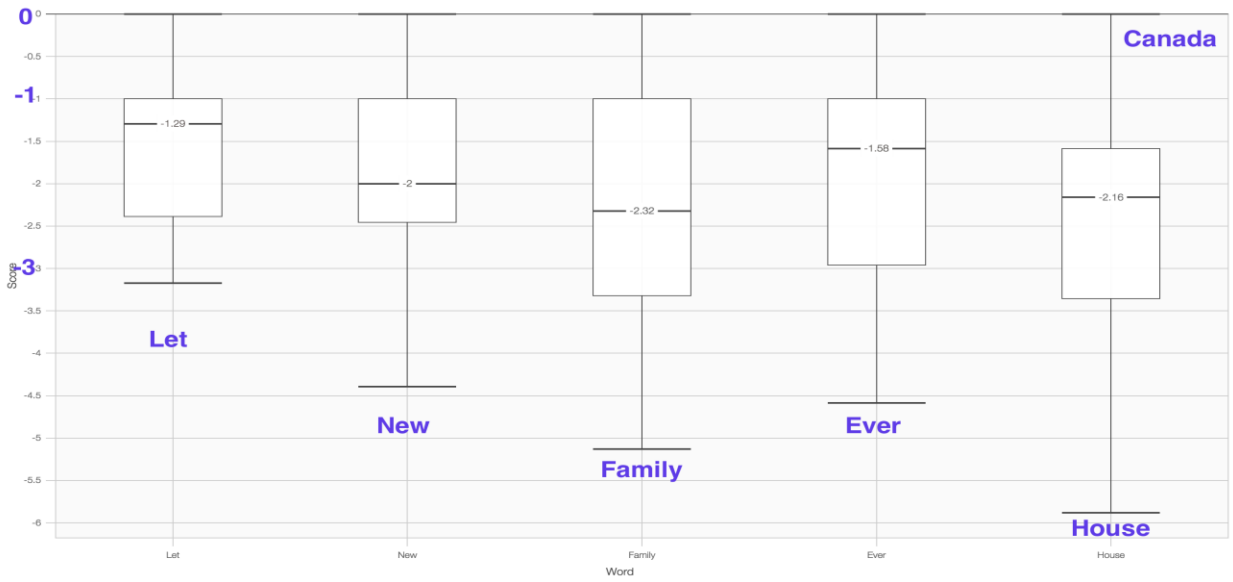

Figure 5: Negative Influence Words

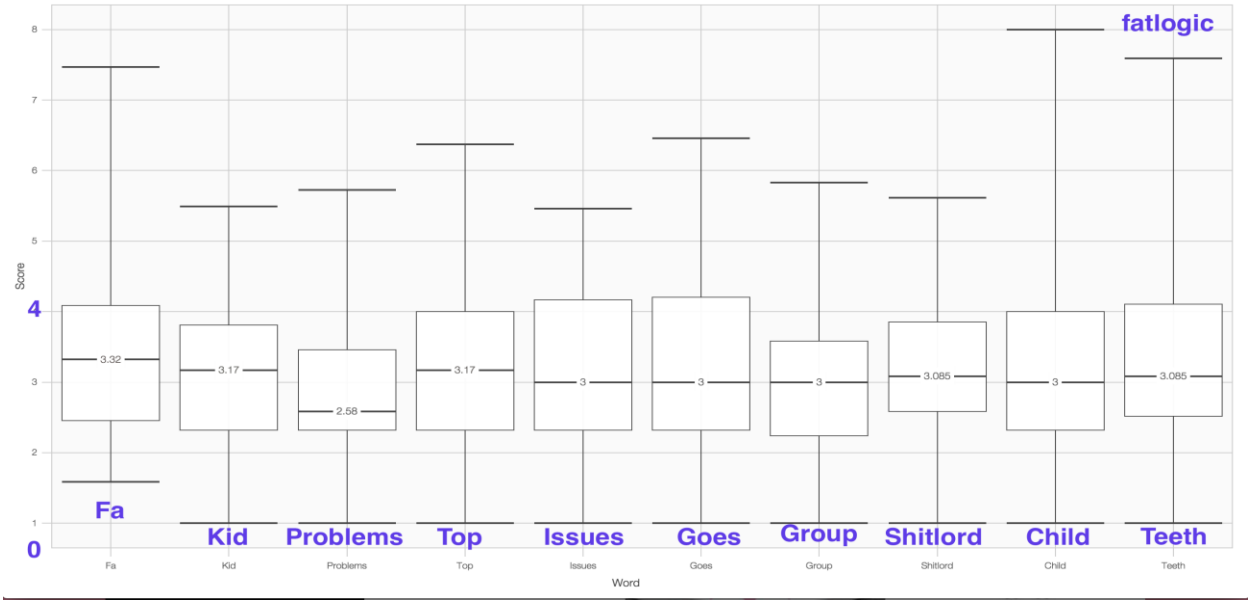

Figure 6: Positive Influence Words

2) From the extracted results, it is possible to identify the special words such as ``fa'' in Figure 6. The definition of ``fa'' is ``Fat Admirer'' and receive a relative higher score in average under thefatlogic subreddit.

3) One finding is that not only the positive sentiment words can obtain a good score but even the negative sentiment words such as ``fucking'' showed in Figure 8. Moreover, there exists some commonly used words but still obtain a high score that is beyond the expectation. For example, the ``yes'' word may reveals the agreement to something so that it is more likely to receive feedback score.

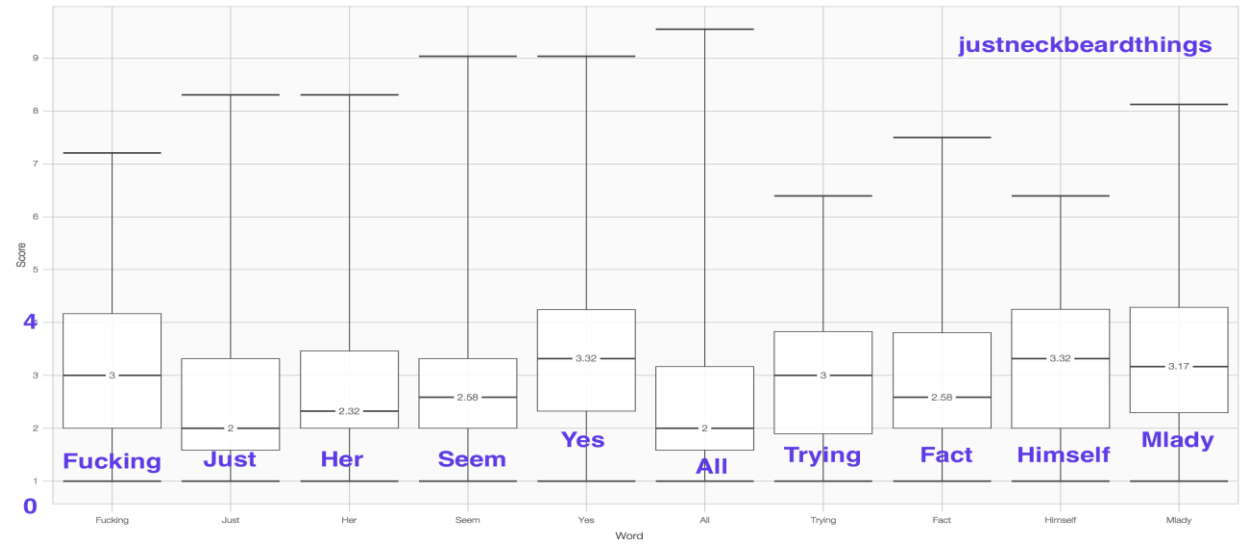

Figure 8: Positive Influence Words

4) Another interesting finding is that there is a word receives high scores across variant unrelated subreddits. We plot the results from 4 randomly selected subreddits in Figure 9. To the best of our knowledge, GT is the abbreviation of "Grand Touring" or “Good Try”. It probably has some other meanings, but it shall depend on the content. We believe this word is extremely useful to earn an up-voting since it is a kind of praise and everybody likes that.

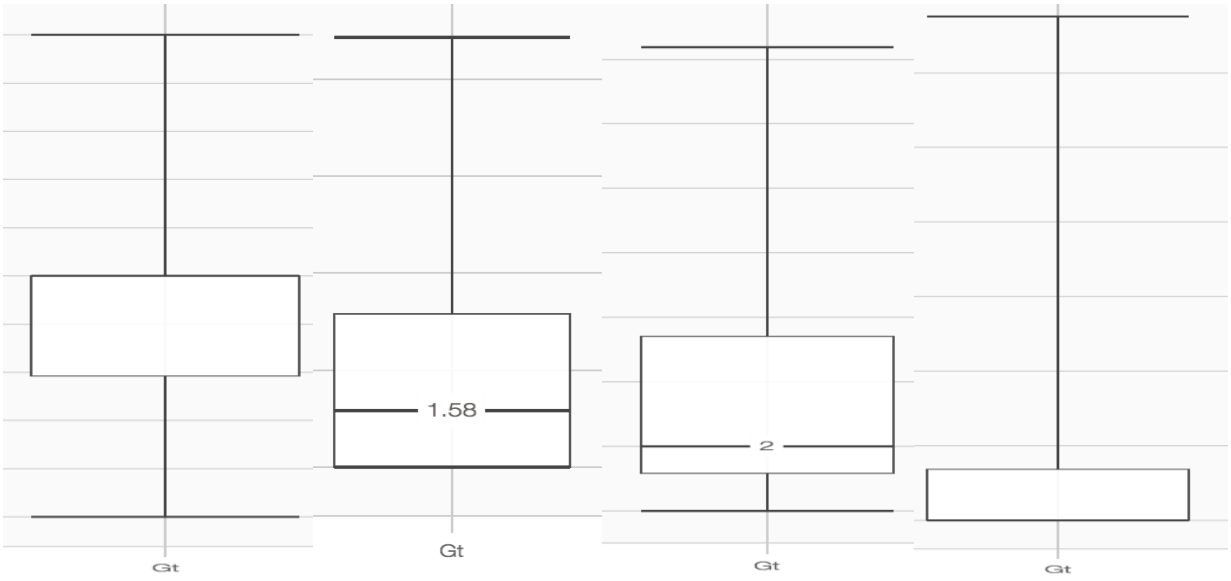

Figure 9: Positive Influence Words

5) Despite the special words we can recognize, there are a lot of high impact words but we do not know the intuitive explanations like the results in Figure 10. It needs a deeper investigation on the original comments.

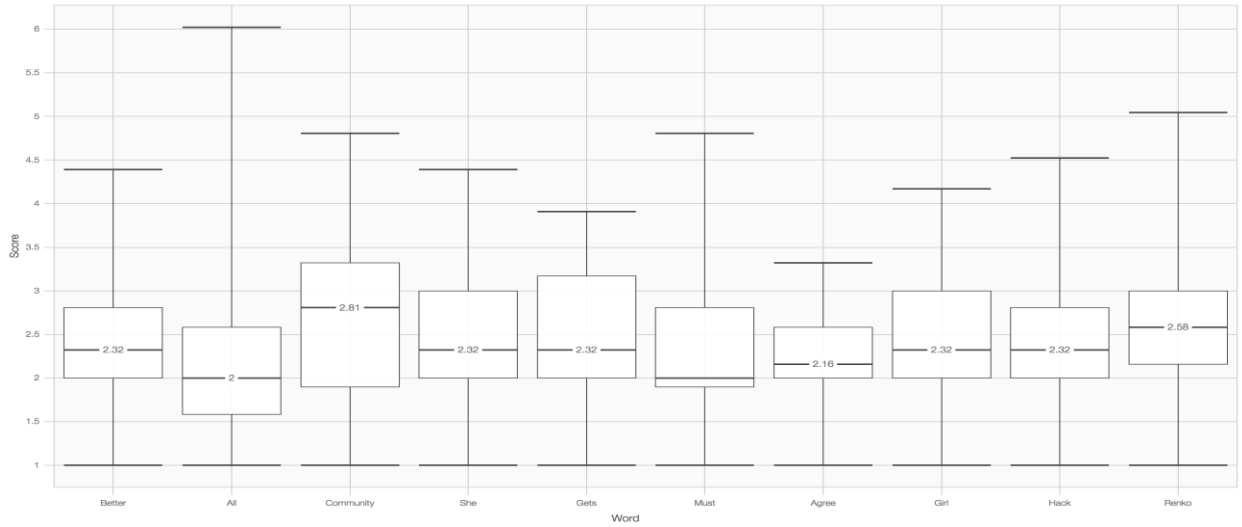

Figure 9: Positive Influence Words