ANALYSIS OF REDDIT GROUPS (SUBREDDITS) USING CLASSIFICATION OF SUBREDDIT POSTS

by

Hira Fatima, BBA, University of Toronto Scarborough, 2010

A Major Research Project
presented to Ryerson University
in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2017

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Hira Fatima

ANALYSIS OF REDDIT GROUPS (SUBREDDITS) USING CLASSIFICATION OF SUBREDDIT POSTS

Hira Fatima

Master of Science 2017

Data Science and Analytics

Ryerson University

ABSTRACT

In this paper, we applied a novel idea to utilize machine learning techniques to automatically label subreddit posts from a subreddit called "askhistorians". Using descriptive analytics, I first conducted an exploratory analysis to see if I can find any patterns, correlations or relationships that could be used to generalize posting pattern and behaviour of reddit users. The second part of my analysis comes from training and evaluating eight classifiers that could correctly categorize reddit posts with a positive or negative label for the eight category codes listed in Appendix A. I used 3 different algorithms and compared their performance using accuracy, precision and recall. This research is a continuation of an existing study that started in Ryerson Social Media Lab (RSML) [1]. The dataset that was used to train and evaluate the classifiers was coded manually by (Ryerson Social Media Lab) RSML. The predicted classification results were used to provide more insights about the subreddit group.

Key words:

Reddit, Subreddits, Text Classification, Linear SVM, Ploy SVM, Naïve Bayes, Social Media

ACKNOWLEDGEMENTS

I am very grateful to **Professor Anatoliy Gruzd** and his team at Ryerson Social Media Lab for their support and assistance in making this project a reality. Professor Anatoliy was my supervisor for this MRP; he has been a great support throughout the term to guide and direct my research and provide valuable feedback. Social Media Lab played a vital role in labelling the dataset manually with classification codes that I used in this project to train my classifiers. It was a manual process that required time and effort and I appreciate the Social Media Lab team completing that task in time for me to use the dataset for this project.

Thank you, Professor Anatoliy and Social Media Lab.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This document provides details on what domain, topic, dataset(s) and research question I chose for my Major Research Project (MRP). I start off by providing a brief background about the topic and datasets, then define the problem and state my research question, followed by literature review and a detailed exploratory analysis of my dataset. I then outline the methodology and experiments performed to train the classifiers, go over the results and close the paper with a discussion about this project and recommendations for future work.

A. Background

For my Major Research Project, I chose to work with Social Media data. There are various kinds of social media platforms that users use for different reason. For example, Facebook is a closed network where its users can choose whom they want to share their posts, comments and data with. On the other hand, platforms such as twitter, a mini blogging site have an open network where anything posted is public. For my project, I will be studying a social news aggregation, web content rating, and discussion website called reddit [1]. As defined on Wikipedia, "Reddit's registered community members can submit content such as text posts or direct links. Registered users can then vote submissions up or down that determines their score and hence, their position on the page. Submissions with the most up-votes appear on the front page or the top of a category. Content entries are organized by areas of interest called "subreddits". Subreddit topics include news, science, gaming, movies, music, books, fitness, food, image-sharing, and many others. The site prohibits harassment, and moderation requires substantial resources." [1] Precisely, for my MRP, my focus and dataset belonged to the subreddit called "askhistorians".

B. Research Question

There are two parts of my analysis, descriptive and predictive analytics. Using descriptive analytics, I wanted to conduct a detailed exploratory analysis of my dataset called "ask-historians" to see if I can find any patterns, correlations or relationships that can be used to generalize posting pattern and behaviour of reddit users. These general patterns or behaviours can then help in evaluating the performance of different groups and maybe even recommend if one group is better at supporting online learning versus the other. The second part of my analysis comes from training and evaluating a classifier that could correctly label reddit posts positive or negative for eight category codes listed in Appendix A. This research is a continuation of an existing study that started in Ryerson Social Media Lab [1]. The dataset that was already coded manually by Ryerson Social Media Lab was used to train and evaluate the classifier. Based on the classification of various posts in each subreddit, the hope was to be able to provide more insights about the subreddit regarding in terms of how well this group supported and promoted online learning.

C. Independent/Dependent Variables

The list of all the available fields in the dataset is given in Appendix C. For the predictive analytics portion of my analysis, ‘Code’ is the output (or dependent) variable. There are 3 independent variables as listed below:

1. ‘Text’: Text classification analysis is used to find key words in the posts (‘text’ field) that could help in coding the post for 8 codes listed in Appendix A.
2. The length of the comment (calculated field)
3. Score of the comment (provided by reddit)

2. LITERATURE REVIEW

The primary focus of this project is to classify subreddit posts automatically leveraging the power of machine learning and then using the results to make inferences about a given subreddit. Although there have been several studies done using reddit data, this automated classification of reddit posts is a novel idea which is implemented in this paper for the first time. In this section, I provide an overview of articles used as a reference for this project.

I started off by reviewing articles related to research on reddit data. There are several studies done using reddit data or on subreddit communities with different objectives; some of which are related to our aim in this project, but none that have implemented the same study before. For instance, Tim Weninger in his article published in 2014 explores the dynamics of hierarchical discussion threads. He conducted an in-depth analysis of subreddit posts on 3 main topics of which the most interesting and related to this project was to find out which variables are the best predictors for high scoring comments [1]. In another article, Nguyen et all studied user’s tweets and extracted information from user tweeting patterns to recommend loosely related Reddit threads [2]. One of the most related articles called “Identifying the social signals that drive online discussions: A case study of Reddit communities” was submitted by Benjamin et all in May 2017. In this article, they conducted sentiment, relevance and content analysis customized for Reddit to develop models that can retrieve top replies under a post with great precision [3]. “Evidence of Online Performance Deterioration in User Sessions on Reddit” is another article published in 2016 that also studied a massive dataset of over 55 million comments posted on Reddit in April 2015. In this article, the authors studies patterns that caused deterioration in the quality of comments produced by users over the course of time [4]. Other studies done on reddit data include “Popularity prediction of Reddit texts” published by Rohlin in 2016 [5] in which the author attempts to predict the popularity of posts on reddit, and “Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?” in which the authors studied reddit’s evolution over a 5 year period of time [6]. Although all these published articles conducted some sort of analysis on Reddit data, my paper goes on to present a novel idea which has not been explored yet.

Since social media is place where individuals can give out their opinions without identifying themselves, there are many times when people use this to their advantage and exhibit undesirable behaviours, one of which is trolling. As defined in Paavola et all’s article called “The Automated Detection of Trolling Bots and Cyborgs and the Analysis of Their Impact in the Social Media” published in 2016, “Trolling is thus a useful tool for any organization willing to force a discussion off-track in the situations when one has no proper facts to back one's arguments.” [7] The paper concludes that sentiment analysis alone cannot detect trolls, however social media analytics tools can be utilized to do this. Based on the same idea, in my project, I also explore the opportunity to identify trolls by applying analytics to my classified posts. For example, if there is a user in my dataset for whom all posts are classified as code C4 “Socializing with Negative Intent”, then we can infer that user to be a troll. In addition to this paper, another article published in May 2014 called “Discovering High-Quality Threaded Discussions in Online Forums” conducted a related machine learning predictive analysis in which the authors tried to predict thread quality without using post rating information [8]. “An Approach to Detect and Analyze the Impact of Biased Information Sources in the Social Media”, is another article that was published in 2015 and studied twitter messages about Ukrainian crisis during 2014 written in Finnish language with the aim to detect trolling behaviour using sentiment analysis [9]. Based on these readings, it is evident that sentiment analysis alone cannot predict trolls, however, in conjunction with social media analytics, it can provide more accurate results. In this paper, our aim is to be able to correctly classify posts in one of eight categories and then use the classification results to narrow down and find users who cause trolling behaviour in online communities.

Before diving deeper into the classification problem, it is vital to explore the dataset and learn as much about it as possible. A detailed exploratory analysis of the dataset was conducted using reference from articles called “Visual Analysis of Social Media Data” [10] written by Tobias Schreck and Daniel Keim in 2013; “Visual analysis of online social media to open up the investigation of stance phenomena” [11] written by Kostiantyn et all in 2015; and, “What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques” written by Annie T Chen1, MSIS, PhD ; Shu-Hong Zhu2, PhD ; Mike Conway2, PhD in 2015 [11]. These papers provided great examples and techniques to use to conduct an exploratory analysis of a text based dataset.

After learning about the dataset, the second step was to solve the classification problem. The classification problem in my project is very much like sentiment analysis. Many of the articles already mentioned earlier including [3], [4] applied sentiment analysis on reddit data to classify posts in different categories. In addition to these articles, another paper that provided a great reference for sentiment analysis was called “Sentiment analysis: A combined approach” published in the Journal of Informetrics in 2013. It provided a great example of combining rule-based classification, supervised learning and machine learning for a new concept called hybrid classification. The authors concluded that this hybrid method improves classification effectiveness and tested their method on movie reviews, product reviews and MySpace comments [12]. These resources provided useful information about how sentiment analysis can be leveraged for classification of posts in my dataset.

Next, I had to figure out which algorithms to use to create my classifier. “The impact of preprocessing on text classification” is a resourceful article that provided details and leads on how to conduct preprocessing on data and which classifier would be optimal. It mentions that SVM is state-of-the-art pattern classifiers and is recommended to be used as the classification algorithm [13]. The best article that I found as a reference similar to what I plan to do in this paper is called “Urdu Text Classification using Majority Voting” published in 2016 in the International Journal of Advanced Computer Science and Applications [14]. This article goes through all the same steps that I intend on using for my text analysis and sentiment classification with explicit examples. The articles used Naïve Bayes, SVM, Random Forest, Bernoulli NB, Multinomial NB algorithms for classification and concluded that Linear SVM provided best results. The authors claimed that they achieved 95% accuracy using maximum voting technique [14]. Using this technique, the authors classified each record using all of the algorithms mentioned and then picked the majority voted class label as the final label. This technique allowed them to further increase the accuracy as opposed to using just one algorithm. Based on the review of these studies, Based on this research, I have decided to use Linear SVM, Ploy SVM and Naïve Bayes classifiers for my post classification.

3. Descriptive Analytics | Exploratory Data Analysis

Since my dataset contains text and numeric fields, I conducted two different kinds of analysis to learn more about my dataset. First, I created a corpus and applied preprocessing techniques including tokenization, stop word and number removal, strip out white spaces, convert to lower case, and stemming to clean the corpus. Using this clean corpus, I conducted further text analysis as outlined below. Note that I did not remove punctuations because based on my preliminary analysis, it was obvious that punctuations are a good predictor of sentiment analysis and it made sense to keep the punctuations for better classification results. The second part of my analysis focused on the numerical fields of the dataset.

*D. **Data Acquisition***

Data for this project is acquired from Reddit using its API. The datasets are open-sourced and compliant with the MRP requirements and have been collected and provided by Ryerson Social Media Lab.

*E. **Data Source & Data Files***

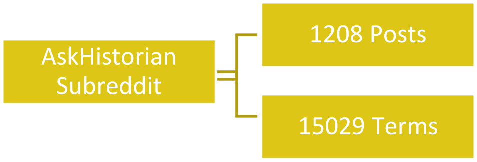
The datasets have been collected by Ryerson Social Media Labb(RSML) for various subreddits (groups) with posts from 2016. For predictive analytics, I used the "askhistorian" dataset that has been coded manually by RSML. This dataset was split into two subsets for training and evaluation of the classifier. Details about the main training/testing dataset are given below:

- The messages were selected randomly from all comments and replies posted in the "askhistorians" subreddit in 2016
- Then they were coded by 3 independent coders. Only scores that were selected by at least 2 out 3 coders were kept
- 19 records were when no two coders agreed were removed from the training dataset (1227 - 19 = 1208).

Data files are posted on GitHub. Link to the GitHub repository and a sample file, both are included in Appendix B.

F. ***Text Analysis***

The dataset contained 1208 posts. When the posts were converted into a document term matrix, the corpus contained 15029 terms in total.



I started off by exploring what kind of text I have in my dataset. The term frequency table is a great way to start because it shows the most frequently used words by their count.

Figure 1 below lists the top 20 words in my corpus. Looking at the words, it seems like most of the posts refer to some link/url because "https" and "www" are in the top word list. "One" is the most frequently used words with 456 occurrences.

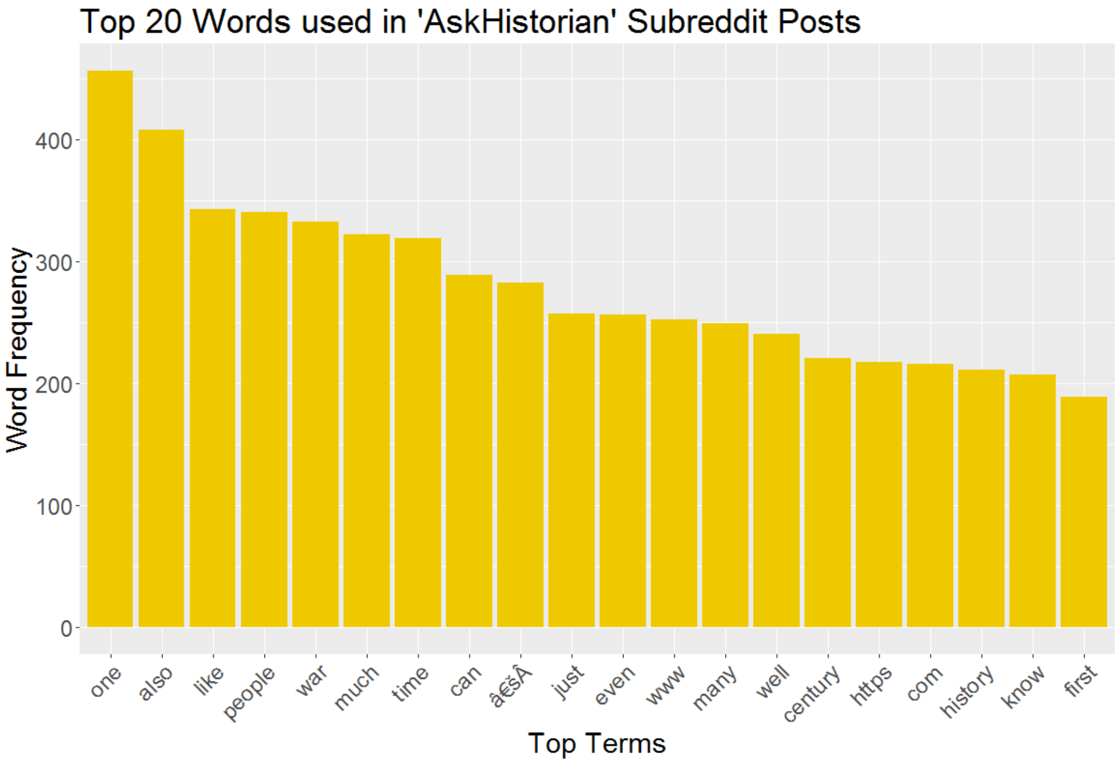


Figure 1 - Top 20 Words used in 'AskHistorian' Subreddit

To get a closer look at the text contained in my dataset, the next visualization I created is a word cloud (Figure 2).

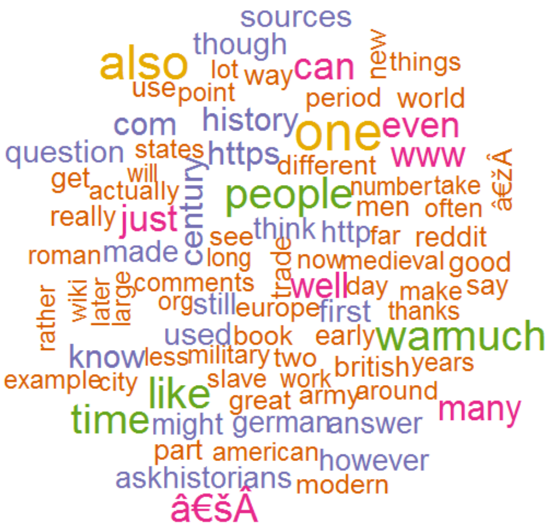


Figure 2 - Word Cloud

The word cloud above lists all words with minimum frequency of 100. Word clouds are useful in understanding what the users are posting about. Most of the words are related to history and it seems like most people posted about European or British history.

After looking at an overview of the text in our corpus, let's move on to do some n-gram analysis. N-grams provide a better context of what the users are posting about as we move to bi and trigrams because these provide most frequent phrases instead of just words. Figure 3 shows that most frequent unigrams are part of a weblink indicating that most of the posts contain some sort of link/url. This finding is in line with what we discovered earlier on using the term frequency chart in Figure 1.

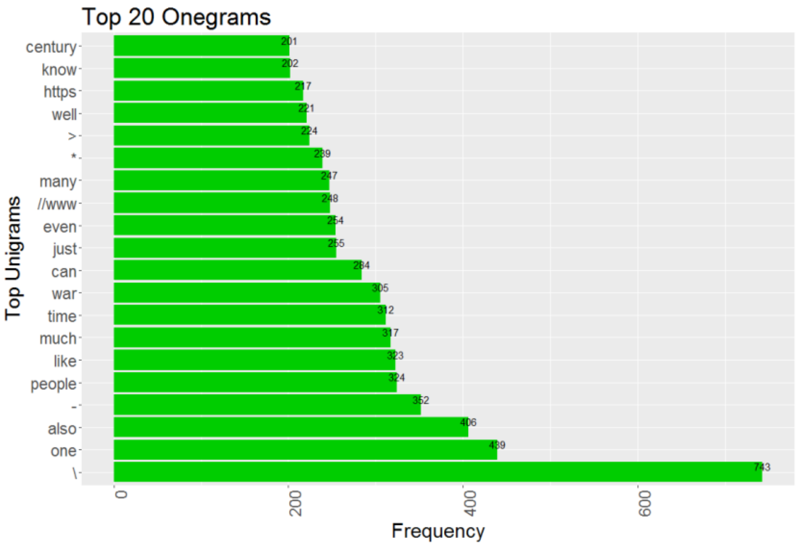


Figure 3 - Top 20 Onegrams

Looking at the bigrams chart in Figure 4, we can confirm that the indication in unigram chart was true and most of the posts do contain some sort of a weblink/url.

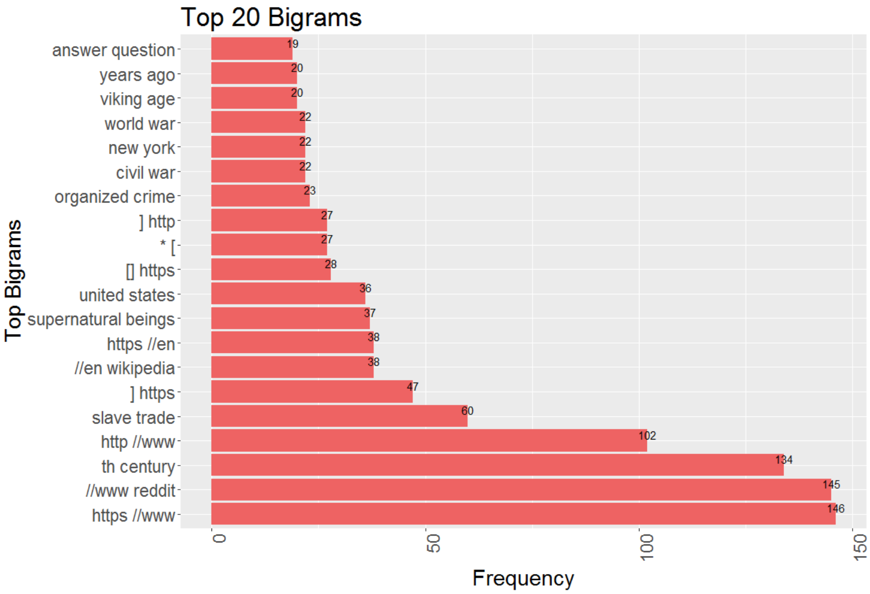


Figure 4 - Top 20 Bigrams

A trigram analysis provides further details on what kind of links are shared in the posts (Figure 5). It seems like most of the links refer to reddit, Wikipedia and reddit rules.

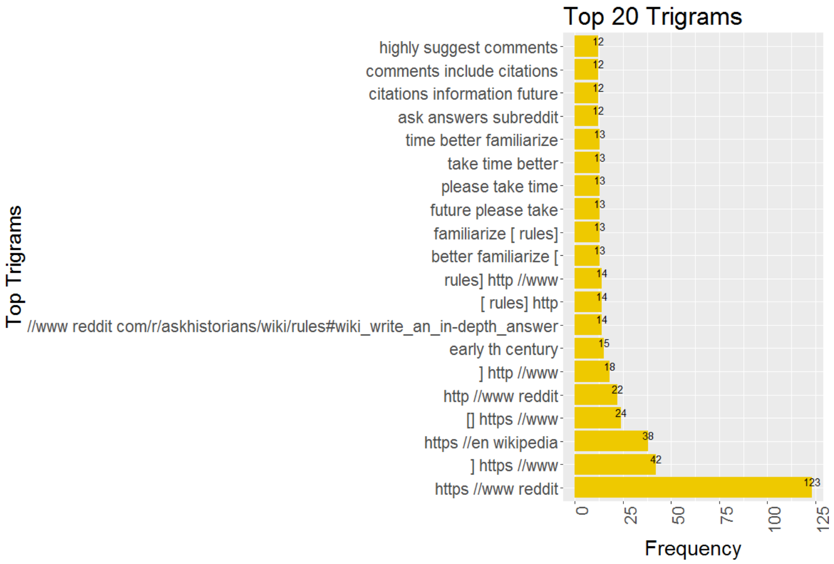


Figure 5 - Top 20 Trigrams

Based on the text analysis, we can conclude that posters often include links/urls in their comments; some of which refer to reddit posts/pages, and others to Wikipedia. Text analysis is important to understand what is contained in our dataset and what the users are posting about. Now that we have a better idea about the text in our dataset, let's move on to data analysis of the numeric fields.

G. *Data Analysis*

The second part of my exploratory analysis focused on numeric fields in the dataset. To perform this analysis, I created a subset of my main dataset with just numeric fields. Let's start off by looking at the word count per post.

Below are summary statistics of 'Word Count per Post' field.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1      25      57     132     137    1605
```

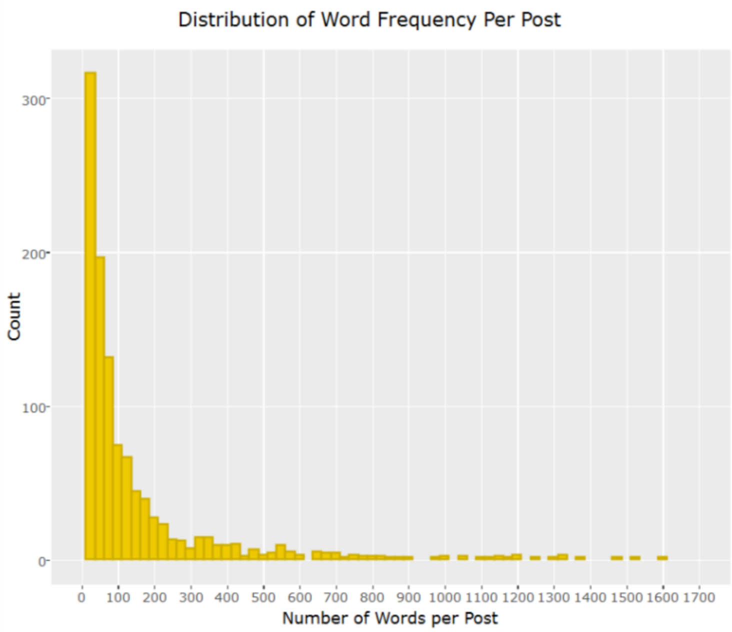


Figure 6 - Distribution of Word Frequency per Post

Based on the summary statistics and Figure 6 above, we can deduce that posts on average have 132 words while accounting for outliers and between 10 and 100 words for majority of the posts excluding outliers. This would mean that users usually write about 3 to 8 sentences approximately per each comment. Next, we move on to study the score distribution.

Below are summary statistics of 'Score per Post' field.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -66.00    1.00    3.00   13.08    8.00 1350.00
```

Since the users can up vote and down vote on posts, the score can be positive or negative. The summary statistics above in addition to Figure 7 show that the minimum score is -66 and many posts have scores greater than 0 and less than 25. This means that the average score including outliers is 13. After looking at the score distribution, next I wanted to understand the distribution of comments versus reply posts.

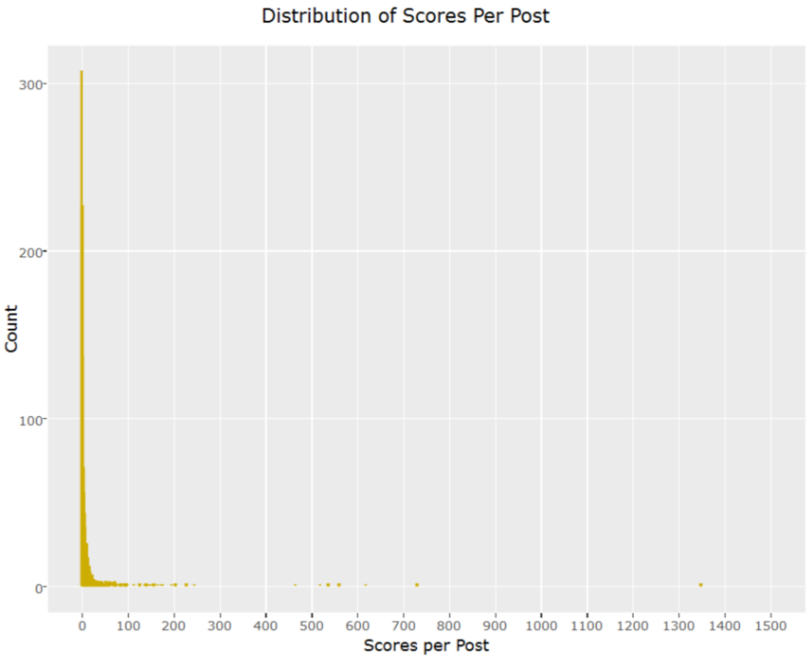


Figure 7 - Distribution of Scores per Post

Figure 8 below shows that most of the posts are reply posts in our dataset. This goes to show that there are conversations happening in this community where users are posting more replies instead of just commenting. Next, we explore the posting pattering of users over the week.

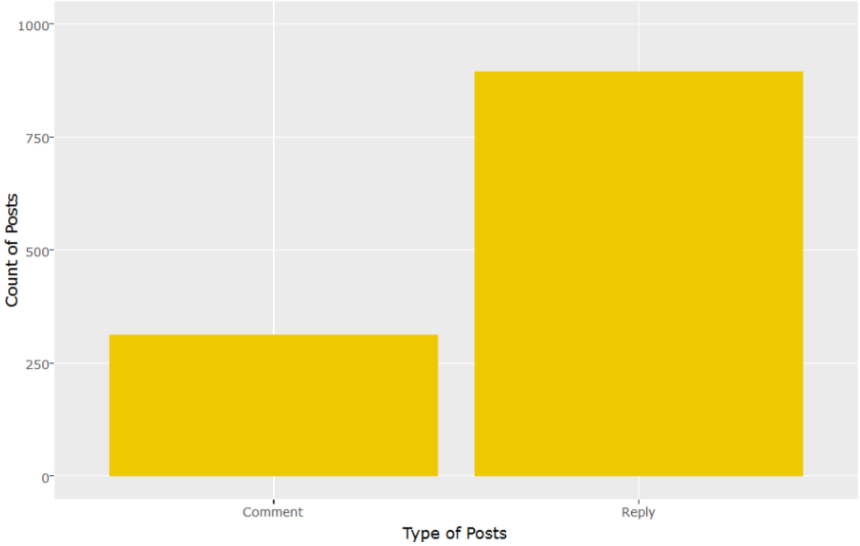


Figure 8- Count of Comments vs Replies

Figure 9 shows posting pattern of users in the ‘askhistorian’ subreddit over the week. The data is split by the type of post. It seems like most of the people reply on Thursdays and Fridays. One anticipated theory to explain this could be that people are more relaxed on Thursdays and Fridays as it is the end of work week and are more inclined to spend time online, hence a spike in the number of posts on Thursdays and Fridays. This theory would need further verification using other subreddit posts data.

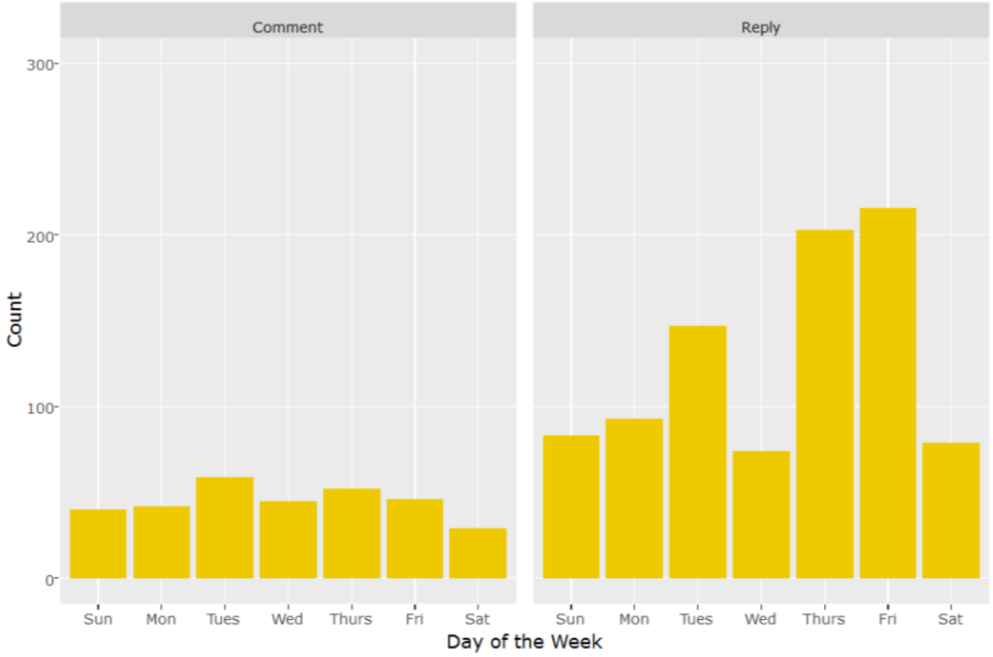


Figure 9 - Posting Pattern over the week

To understand and study the overall health of any given subreddit, the dataset needs to be classified for the eight codes. Figure 10 shows code distribution of ‘askhistorian’ dataset that has already been manually classified. Majority of the posts were classified as C3 which stands for “Explanation with Neutral Presentation”. This means that most of the users’ posts included explanation with a neutral tone. The second most popular code is C6 “Information Seeking” followed by C7 “Providing Resources”. Dominance of these codes shows that the subreddit promotes sharing of information.

The fact that code C4, “Socializing with Negative Intent” has the least number of associated posts confirms that this subreddit is not a victim of trolling as there are very few posts classified as such. We can also infer that the moderators are doing a good job in keeping trolls away from this group.

Note that there is some overlap in coding as well because one post can be positive for more than one codes, and it is for this reason that I created separate binary classifiers for each code in the machine learning experiment.

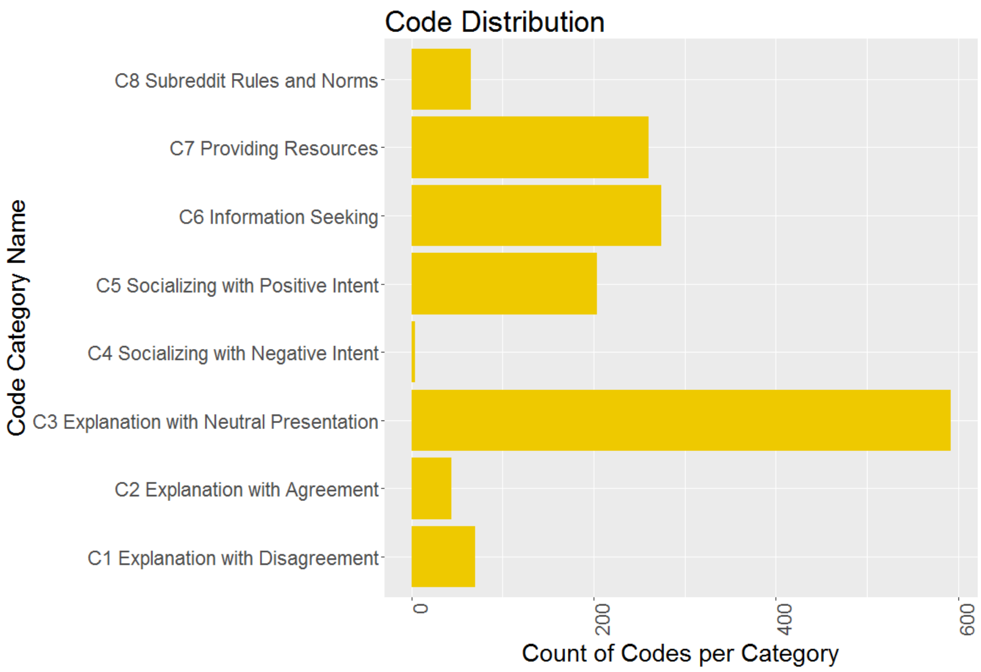


Figure 10 - Code Distribution

Figure 11 shows the distribution of posts in terms of the depth of thread. Majority of the posts have depth less than 10 which means that the posts have 10 or less comments and/or replies per question asked.

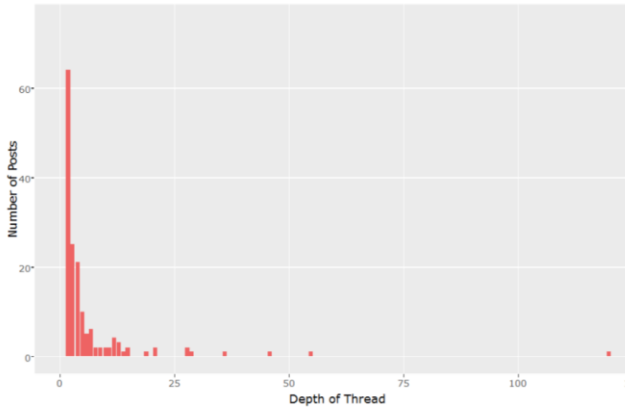
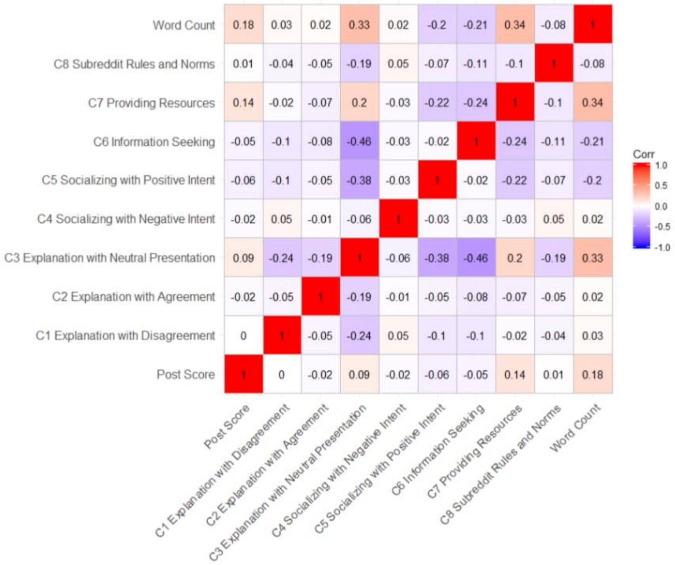


Figure 11 - Distribution of posts in terms of Depth of Thread



	Post Score	C1 Explanation with Disagreement	C2 Explanation with Agreement	C3 Explanation with Neutral Presentation	C4 Socializing with Negative Intent	C5 Socializing with Positive Intent	C6 Information Seeking	C7 Providing Resources	C8 Subreddit Rules and Norms	Word Count
Word Count	0.18	0.03	0.02	0.33	0.02	-0.2	-0.21	0.34	-0.08	1
C8 Subreddit Rules and Norms	0.01	-0.04	-0.05	-0.19	0.05	-0.07	-0.11	-0.1	1	-0.08
C7 Providing Resources	0.14	-0.02	-0.07	0.2	-0.03	-0.22	-0.24	1	-0.1	0.34
C6 Information Seeking	-0.05	-0.1	-0.08	-0.46	-0.03	-0.02	1	-0.24	-0.11	-0.21
C5 Socializing with Positive Intent	-0.06	-0.1	-0.05	-0.38	-0.03	1	-0.02	-0.22	-0.07	-0.2
C4 Socializing with Negative Intent	-0.02	0.05	-0.01	-0.06	1	-0.03	-0.03	-0.03	0.05	0.02
C3 Explanation with Neutral Presentation	0.09	-0.24	-0.19	1	-0.06	-0.38	-0.46	0.2	-0.19	0.33
C2 Explanation with Agreement	-0.02	-0.05	1	-0.19	-0.01	-0.05	-0.08	-0.07	-0.05	0.02
C1 Explanation with Disagreement	0	1	-0.05	-0.24	0.05	-0.1	-0.1	-0.02	-0.04	0.03
Post Score	1	0	-0.02	0.09	-0.02	-0.06	-0.05	0.14	0.01	0.18

Figure 12 - Correlation Matrix for Numerical Fields in the Dataset

Figure 12 here provides a correlation matrix of the numerical fields in the dataset. There are no significant relationships shown in this matrix; however, some of the weaker relations do make sense and can be explained. For instance, C3 (Explanation with Neutral Presentation) is negatively correlated with C6 (Information Seeking). It makes sense that a user would either be providing explanation if they are knowledgeable or asking for information if they do not know about the subject; hence the two are opposite of each other. This is somewhat evident with a weak negative correlation. Similarly, C7 (Providing Resources) and C3 (Explanation with Neutral Presentation) have weak positive relationships with Word Count. This also makes sense because if someone is providing resources or explanation, they would be using more words and therefore the word count of the post would be higher.

H. Extended Exploratory Analysis Using Tableau

Here is the link to my Tableau Workbook where I performed more detailed exploratory analysis. https://public.tableau.com/views/MRPExploratoryAnalysis/AvgScorePostCountperUser?:embed=y&:display_count=yes&publish=yes

The first screenshot (Figure 13) below shows the first chart that contains code distribution by authors. This interactive chart can be sorted for all codes to study posting pattern of different users and identify trolling users. In this current dataset, we only have 4 instances where a post was classified as C4, once by each of the 4 users. Since we have a very sparse set of records for C4, trolling behaviour and users cannot be identified in this example. But after we apply the same metrics on other subreddits, and, if we can identify users for whom all posts are classified as C4, then we can infer that they are possibly trolls. This hypothesis would require further study.

In Figure 13 below, we can note that "itsallfolklore" is one of the most active users with posts classifications in almost all categories except C4.

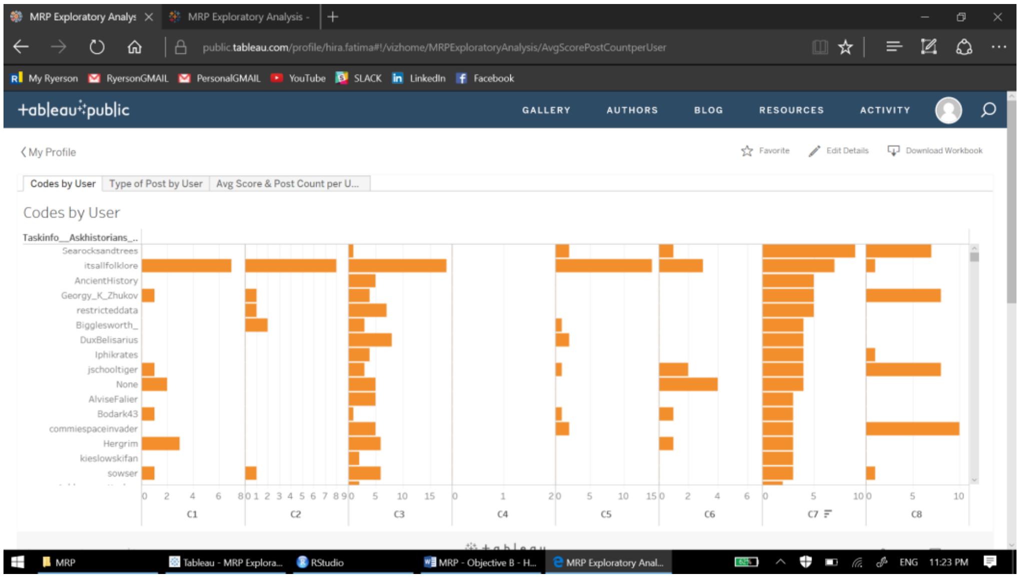
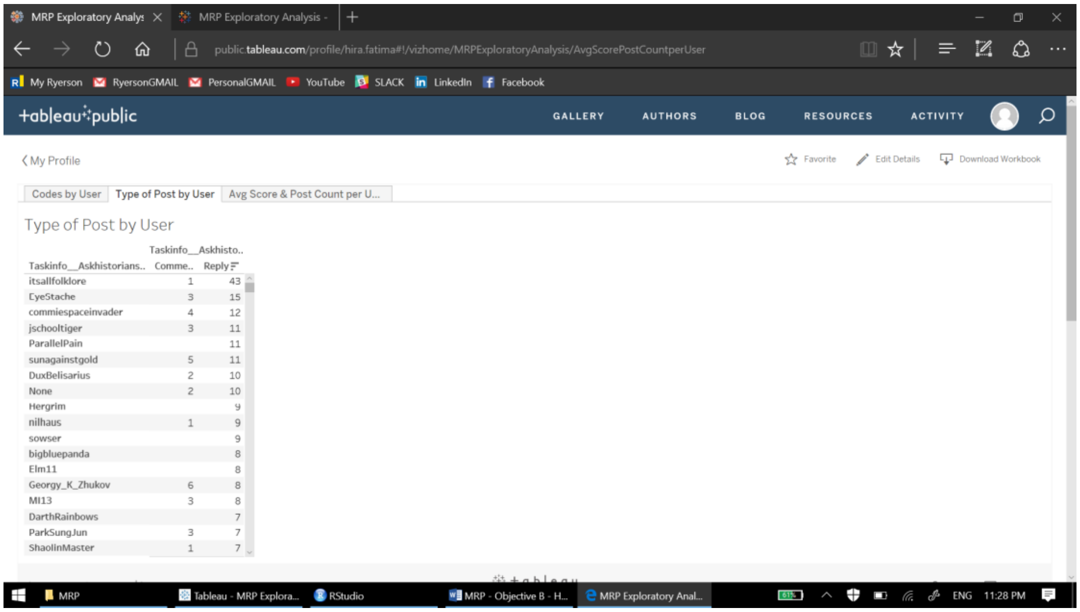


Figure 13- Most Active Users (Tableau View)

Figure 14 shows a distribution of the number of posts split by type (comment vs. reply). This chart shows that author "itsallfolklore" is the most involved user in this subreddit with the highest number of reply posts.



Type of Post by User

	Taskinfo__Askhisto..	
Taskinfo__Askhistorians..	Comme..	Reply
itsallfolklore	1	43
EyeStache	3	15
commiespaceinvader	4	12
jschooltiger	3	11
ParallelPain		11
sunagainstgold	5	11
DuxBelisarius	2	10
None	2	10
Hergrim		9
nilhaus	1	9
sowser		9
bigbluepanda		8
Elm11		8
Georgy_K_Zhukov	6	8
MI13	3	8
DarthRainbows		7
ParkSungJun	3	7
ShaolinMaster	1	7

Figure 14 - Comments versus Replies distribution (Tableau View)

Looking at the Figure 15 below, it is obvious that "itsallfolklore" is the most popular and dominant user in this subreddit. This user has the highest average score for their posts and the highest number of comments/replies posted in this subreddit.

If we could use this information to recommend users as moderators for a given subreddit, then using this example, "itsallfolklore" could be a potential recommendation as a moderator.

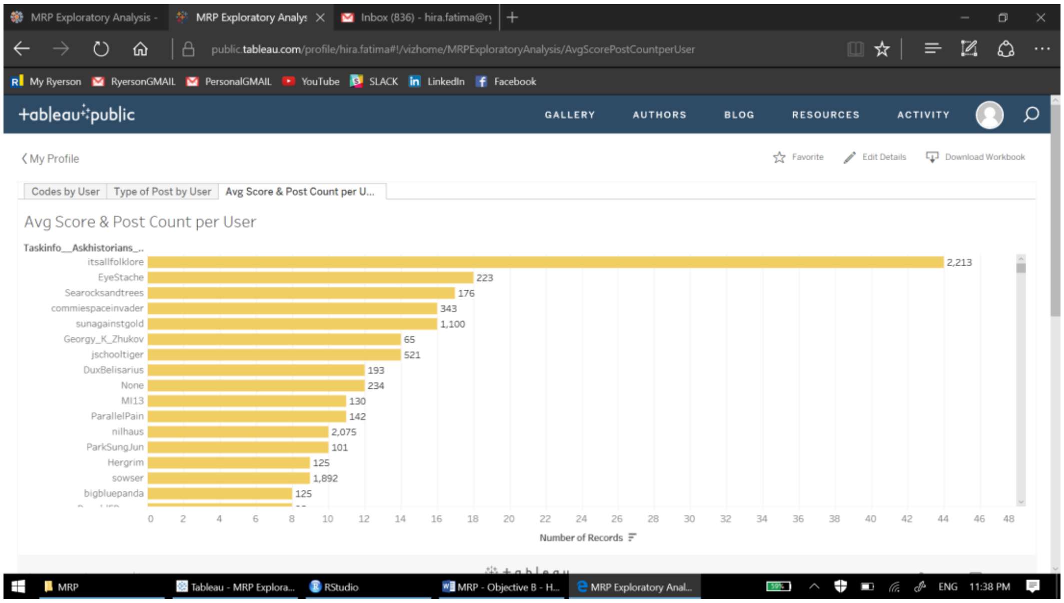


Figure 15 - Average Score and Post Count per User (Tableau View)

I. *Factor Analysis*

Factor Analysis was also conducted to understand how the different codes were related or not as part of the Exploratory Data Analysis. Two iterations of factor analysis were conducted, the first one using R Programming and the second one using SPSS. Figure 16, 17, and 18 show the results of Factor Analysis conducted in R. For this iteration, Word Count and Post Score were also added to the factor analysis to understand if there was any correlation amongst these variables along with the classification codes (C1 to C8).

Figure 16 shows that none of the variables are positively correlated to each other. There is a weak negative correlation between C7 and C6 as well as C7 and C5.

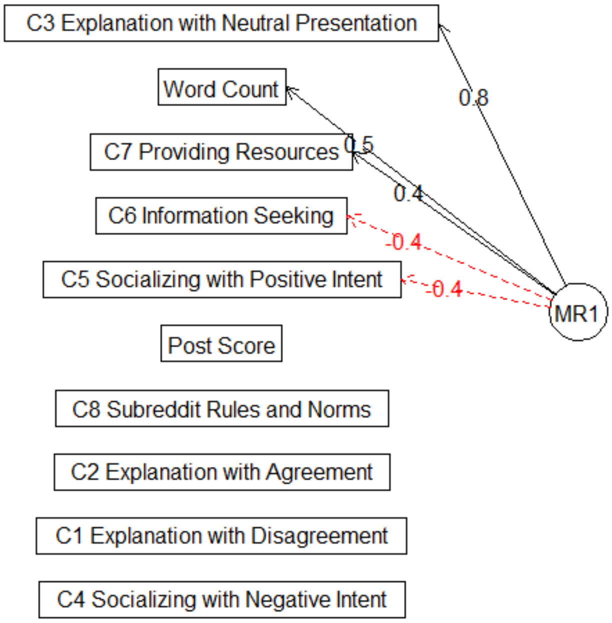
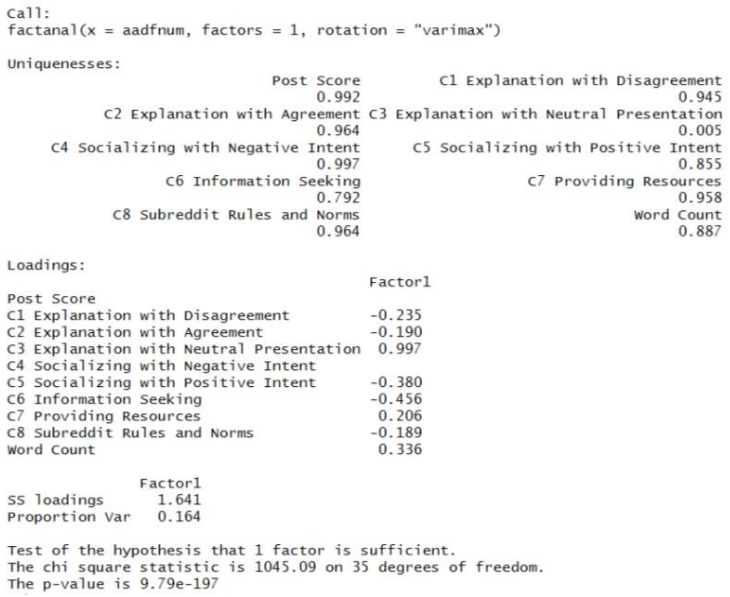

C3 Explanation with Neutral Presentation
Word Count
C7 Providing Resources
C6 Information Seeking
C5 Socializing with Positive Intent
Post Score
C8 Subreddit Rules and Norms
C2 Explanation with Agreement
C1 Explanation with Disagreement
C4 Socializing with Negative Intent
0.8
0.5
0.4
-0.4
-0.4
MR1

Figure 16 - Factor Analysis

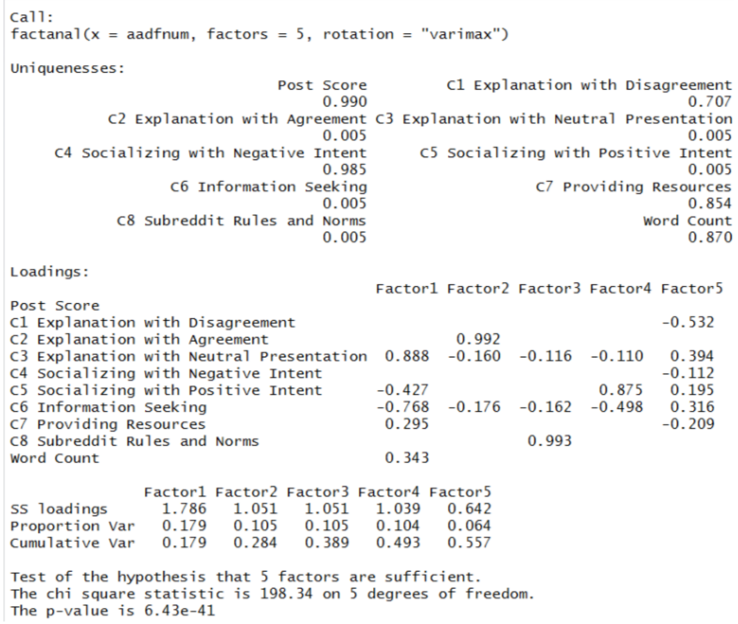
Factor analysis was repeated a few times using a different number of factors in the output criteria. Figure 17 and Figure 18 show the results for 1 and 5 factors, respectively. The output of the result shows that the value of C3 is the highest when 1 factor is used.



```
Call:
factanal(x = aadfnum, factors = 1, rotation = "varimax")

Uniquenesses:
                              Post Score        C1 Explanation with Disagreement
                                   0.992                                   0.945
          C2 Explanation with Agreement C3 Explanation with Neutral Presentation
                                   0.964                                   0.005
     C4 Socializing with Negative Intent      C5 Socializing with Positive Intent
                                   0.997                                   0.855
                   C6 Information Seeking                  C7 Providing Resources
                                   0.792                                   0.958
            C8 Subreddit Rules and Norms                              Word Count
                                   0.964                                   0.887

Loadings:
                                         Factor1
Post Score
C1 Explanation with Disagreement         -0.235
C2 Explanation with Agreement            -0.190
C3 Explanation with Neutral Presentation  0.997
C4 Socializing with Negative Intent
C5 Socializing with Positive Intent      -0.380
C6 Information Seeking                   -0.456
C7 Providing Resources                    0.206
C8 Subreddit Rules and Norms             -0.189
Word Count                                0.336

               Factor1
SS loadings      1.641
Proportion Var   0.164

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 1045.09 on 35 degrees of freedom.
The p-value is 9.79e-197
```

Figure 17 - Output with 1 Factor

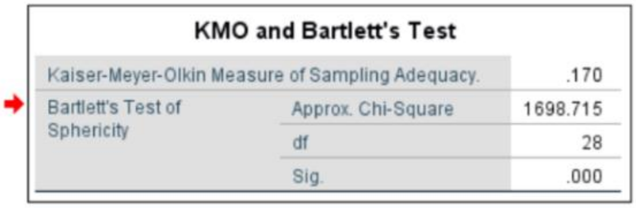
Figure 18 shows that when 5 factors were used, C3 has the highest loading on Factor 1, C2 on Factor 2, C8 on Factor 3, and C5 on Factor 4. There is no instance where multiple codes have a significantly high loading for any one factor. This confirms that the questions are not likely to be correlated.



```
Call:
factanal(x = aadfnum, factors = 5, rotation = "varimax")

Uniquenesses:
                              Post Score        C1 Explanation with Disagreement
                                   0.990                                   0.707
          C2 Explanation with Agreement C3 Explanation with Neutral Presentation
                                   0.005                                   0.005
     C4 Socializing with Negative Intent      C5 Socializing with Positive Intent
                                   0.985                                   0.005
                   C6 Information Seeking                  C7 Providing Resources
                                   0.005                                   0.854
            C8 Subreddit Rules and Norms                              Word Count
                                   0.005                                   0.870

Loadings:
                                         Factor1 Factor2 Factor3 Factor4 Factor5
Post Score
C1 Explanation with Disagreement                                         -0.532
C2 Explanation with Agreement                     0.992
C3 Explanation with Neutral Presentation  0.888  -0.160  -0.116  -0.110   0.394
C4 Socializing with Negative Intent                                      -0.112
C5 Socializing with Positive Intent      -0.427                   0.875   0.195
C6 Information Seeking                   -0.768  -0.176  -0.162  -0.498   0.316
C7 Providing Resources                    0.295                          -0.209
C8 Subreddit Rules and Norms                              0.993
Word Count                                0.343

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      1.786   1.051   1.051   1.039   0.642
Proportion Var   0.179   0.105   0.105   0.104   0.064
Cumulative Var   0.179   0.284   0.389   0.493   0.557

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 198.34 on 5 degrees of freedom.
The p-value is 6.43e-41
```

Figure 18 - Output with 5 Factors

In addition to the above factor analysis, an additional iteration conducted using SPSS confirmed that it was not advisable to conduct Factor Analysis on our given dataset as per the results shown below. The results (Figure 19) from the Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test of Sphericity were below satisfactory levels to proceed with Factor Analysis. KMO value came as 0.17 which is much lower than the advised minimum value of 0.6.



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.170
Bartlett's Test of Sphericity	Approx. Chi-Square	1698.715
	df	28
	Sig.	.000

Figure 19 - KMO Test Results

4. METHODOLOGY AND EXPERIMENTS

*J. **Aim of Study***

The aim of this study is to train binary text classifiers that could automatically assess if a given subreddit post belongs to Code 1 to 8 listed in Appendix – A | Reddit Codebook. A different binary text classifier is trained for each code as one post could belong to more than 1 code. To determine the best trained classifier, 3 different algorithms were used and compared statistically to determine the best fit for this test.

*K. **Response (Dependent) and Independent Variable(s)***

In our experiment, the response variables are C1, C2, C3, C4, C5, C6, C7, and C8. Although the dataset contained more fields, only the following were used as inputs to train the classifier.

- Score
- Word Count (Calculated Field)
- Text (converted into Term Document Matrix)

*L. **Factors and Levels***

In our experiment, the factors are the different algorithms being tested. Since we are using 10-fold cross validation, the classifier automatically picks the optimal levels for each algorithm.

M. ***Experimental Design***

a) Text Processing

The dataset called ‘askhistorian’ contained both numeric and text fields. In the first step, the text field which contained all subreddit posts was processed and converted into a Term-Document Matrix to be then used as input for the classifier. The text is pre-processed to remove any numbers, strip out any white spaces, covert all text to lower case and remove stop words. Different combinations of pre-processing were tried and finally it was determined to keep the punctuations because with punctuations, the results were slightly better.

b) Randomization (Train/Test Split)

The dataset was divided randomly into two sets, one for training with 70% data, and another for testing with 30% data.

c) Cross Validation

Since the dataset was not very large, it was important to implement a validation technique. In our experiment, I used 10-fold cross validation method. As a result, the classifiers were also able to pick the best parameter settings that provided optimal results.

N. ***Experiment Performance and Revisions***

A series of experiments were conducted to fine tune the classifiers and improve model accuracy, precision and recall.

Details of different experiments are provided below.

a) Experiment 1

In the first experiment, I trained 3 classifiers using SVM Linear, SVM Poly, and Naïve Bayes algorithms for each of the 8 category codes. Without making any changes to the model, I altered text processing parameters to test which combination of pre-processing steps were yielding results with highest accuracy. Once that was determined, I then moved on to evaluate classifier performance.

b) Experiment 2

In the second experiment, my primary focus was to understand how the classifiers were performing in terms of accuracy, precision and recall. Although Naïve Bayes provided highest accuracy for majority of the codes, the precision and recall were very poor. I could figure this out only after I examined the confusion matrices for each code one by one. Under Naïve Bayes, the accuracy was high because the dataset was highly imbalanced and almost all posts were being classified with the majority class (negative label). As a result, the precision and recall were very low. Consequently, I decided to focus on precision and recall measures to evaluate classifier performance instead of accuracy alone. At this stage, codes C1, C2, C4 and C8 were also dropped from the experiment because of very low ratio of positive class that resulted in poor precision and accuracy. Only codes with at least 10% instance of positive class label (or minimum 100 records) were included in the next experiment.

c) Experiment 3

In the third experiment, I focused on finding optimal parameters for imbalanced datasets. Using a great blog called "Practical Guide to deal with Imbalanced Classification Problems in R" I implemented the different techniques used to deal with imbalanced data including Undersampling, Oversampling, Synthetic Data Generation, and Cost Sensitive Learning [15]. Following the example in the blog, I implemented all the different method and the best results came from undersampling (downsampling) technique.

*O. **Measuring Classifier Performance***

Precision and Recall were the main measures used to evaluate classifier performance. Although the accuracy was also looked at, precision was a more significant measure because the dataset was imbalanced and a high accuracy when everything gets assigned to the majority class is misleading.

*P. **Algorithm Comparison and Selection***

To compare the algorithms, statistical summary was calculated for each algorithm based on the results obtained for each code. The algorithm with optimal mean, median, max and min was selected for further study and as the final recommended model to use.

5. RESULTS AND DISCUSSION

*Q. **Exploratory Analysis Results***

The exploratory analysis of this subreddit data provided some useful leads for further exploration of subreddit data for specific purposes. For instance, using the results from code classification such as C7 (Providing Resources) along with other information such as the number of posts over time and level of activity, we can recommend moderators for a given subreddit. A user whose posts are mostly classified positive for C7 (Providing Resources) and also has a significantly large number of replies and comments can be flagged as being a potential candidate to be the moderator of the subreddit.

As another example, the exploratory analysis conducted along with the features of our dataset confirmed that the forum is well regulated and the moderators are doing a decent job with keeping out trolling activity as we did not have enough examples of C4 code in our dataset. In the event that trolling was happening in our test subreddit, we would have certainly expected to see more examples of posts labelled positive for code C4.

Similarly, the exploratory factor analysis also helped in confirming that the codes were well designed and although the same post can be classified with more than 1 code, the codes themselves are not correlated. This helps to confirm that each of the 8 codes has its own value and there is no need to merge any of these together.

Aside from providing more details about the codes and their classification, this analysis also provided more insight into the data and what kind of messages users have been posting in this subreddit. It is important to understand that this subreddit revolved around the topic of history as the name suggests. However, if we were to look at a different subreddit, the text distribution can be much more different. Since the dataset is so specific to a given interest line (history in this instance), it cannot necessarily be generalized for other subreddits. In order to apply this same study to other subreddits, it is recommended to retrain a dataset with messages from that particular domain for better results.

R. ***Machine Learning Experiment Results***

A series of experiments was conducted to find out the optimal solution using different approaches and techniques as outlined below:

a) Experiment 1

As part of the first experiment, my focus was on figuring out which combination of text processing parameter would yield optimal results for any given classifier. After testing different combinations, it was concluded that punctuations did make a difference in providing better accuracy. This made sense because if someone is asking a question, they are bound to use a question mark or users who want to express some sentiment in their post would be highly likely to use an exclamation mark.

b) Experiment 2

In the second experiment, codes C1, C2, C4 and C8 were dropped from the analysis due to lack to sufficient representative examples.

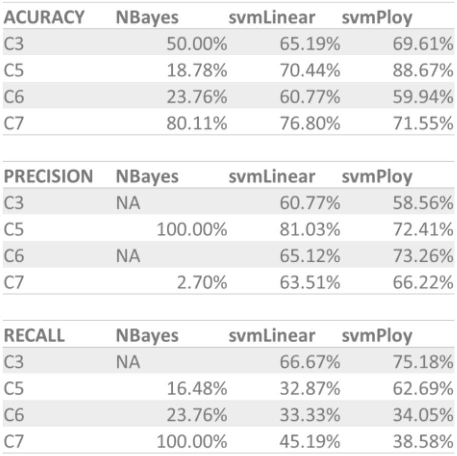
While I was analyzing the confusion matrices one by one, I noticed that Naïve Bayes was giving much higher accuracies but the confusion matrix didn't look very promising as shown below.

```
##        0     1
##   0  288     0
##   1   72     2
```

In the example above, majority of the records were classified as 0 when they were truly 0. And only 2 of the records were classified as 1 when they were truly 1. There were 72 incorrect classifications; however, since 288 was a majority count, it outweighed the incorrect classifications resulting in an accuracy of 80.11%.

With this information, it became evident to work with recall and precision as opposed to accuracy as it could be misleading given the imbalanced nature of our dataset.

Table 1 shows the precision and recall for these classifiers. Although the values were not high enough as I was hoping, there was still one more thing I could try to improve the results; and that was applying techniques for imbalanced datasets.

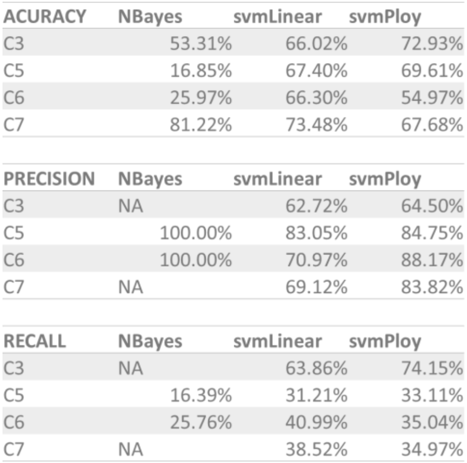


ACURACY	NBayes	svmLinear	svmPloy
C3	50.00%	65.19%	69.61%
C5	18.78%	70.44%	88.67%
C6	23.76%	60.77%	59.94%
C7	80.11%	76.80%	71.55%

PRECISION	NBayes	svmLinear	svmPloy
C3	NA	60.77%	58.56%
C5	100.00%	81.03%	72.41%
C6	NA	65.12%	73.26%
C7	2.70%	63.51%	66.22%

RECALL	NBayes	svmLinear	svmPloy
C3	NA	66.67%	75.18%
C5	16.48%	32.87%	62.69%
C6	23.76%	33.33%	34.05%
C7	100.00%	45.19%	38.58%

Table 1 - Without applying techniques for imbalanced dataset



ACURACY	NBayes	svmLinear	svmPloy
C3	53.31%	66.02%	72.93%
C5	16.85%	67.40%	69.61%
C6	25.97%	66.30%	54.97%
C7	81.22%	73.48%	67.68%

PRECISION	NBayes	svmLinear	svmPloy
C3	NA	62.72%	64.50%
C5	100.00%	83.05%	84.75%
C6	100.00%	70.97%	88.17%
C7	NA	69.12%	83.82%

RECALL	NBayes	svmLinear	svmPloy
C3	NA	63.86%	74.15%
C5	16.39%	31.21%	33.11%
C6	25.76%	40.99%	35.04%
C7	NA	38.52%	34.97%

Table 2 - After applying techniques for imbalanced dataset

c) Experiment 3

The final results after applying the optimal technique to work with imbalanced dataset are given in Table 2. Undersampling technique for imbalanced datasets provided optimal results for our dataset. Looking at the Naïve Bayes results, it is clear that the accuracy is fairly high; however, the precision and recall are not favourable. For this reason, it was concluded that Naïve Bayes performed the worst of all algorithms and was not suitable for this experiment as our dataset was highly imbalanced.

As far as a comparison between SVM Linear and Poly is concerned, SVM Linear generally performed better for all classifiers; however, it is vital to note that it is not necessary that the same algorithm should work optimally for all binary classifiers. In our example, for C7, SVM Ploy did better based on the precision, however, when we take recall into account, SVM Linear provideed better results. For C5, based on precision and recall, SVM Ploy performed better as opposed to SVM Linear.

S. ***Discussion***

This project put to practice a novel idea of applying automated text classification to subreddit posts. Based on the code labels, insights could be extracted out of text data. Although efforts were made to try different techniques to optimize prediction accuracy and recall, the results were limited due to some challenges faced. The first and foremost challenge was around limitation of training dataset. Although the dataset contained over 1000 records, the distribution of examples with positive label for a given code was very imbalanced. In fact, 4 out 8 codes had less than 100 positive examples (which is less than 10% of the dataset) because of which these codes were excluded from the experiment in the last stage. Secondly, having a bigger dataset would have certainly helped in increasing the accuracy overall; however, in our case our dataset was limited to 1208 records which can be argued as being an inadequate size of dataset for this kind of a classification task.

Figure 20 below shows the ROC curves generate for the SVM Linear Classifier for codes C3, C5, C6 and C7. It shows that the AUC (Area under the Curve) is mostly in the 60% to 76% range for all these classifiers. Although we would have liked to get a higher value for AUC, based on the limitations mentioned in the previous paragraph, it was not possible in this iteration of experiments.

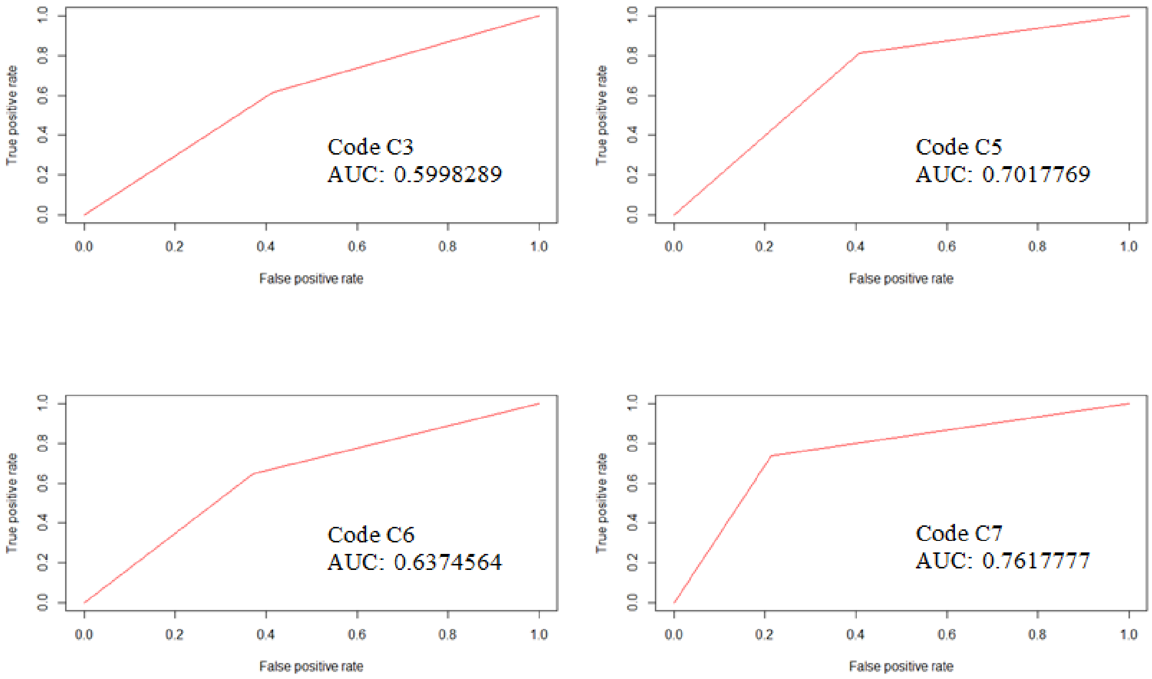


Figure 20 – ROC Curves and AUC for SVM Linear Model

To apply this same concept and classification to other subreddits, it would be beneficial to have more training data with example from the same subreddit (as that being classified). This is because the ‘askhistorian’ subreddit contains a specific set of words and the classifiers are based on this particular bag of words (tokens). When transitioning to another subreddit, the bag of words can be different enough to significantly impact the classifier’s ability to predict correctly. This can be a suggestion for future work to evaluate if a trained classifier for one subreddit would perform any better, same or worse on another subreddit's posts. Also, people in different subreddits many have a different style of writing and that can also influence the way the classifier is trained especially because we had decided to keep the punctuation as a positive predictor of sentiment.

6. CONCLUSION AND FUTURE WORKS

For further exploration of this study, it would be a great idea to repeat the process for another subreddit and compare the two to see if there is any improvement in the results. Since the key words in each subreddit can be vastly different based on the domain, using the same classifier across different subreddits may not reveal optimal results. This can only be confirmed with further study.

Another great finding of this project was the reassurance that Naïve Bayes is not a good and appropriate contender when testing algorithms on imbalanced datasets. Keeping this in mind, perhaps it would be better to test a different algorithm other than Naïve Bayes.

Our dataset did not contain enough examples of records with code C4 (Socializing with Negative Intent), however if it did, and if most or all of the posts were associated with the same user account, then that user could be classified as a troll. This is a potential topic for further study and can also be broadened across various subreddits. For instance, if one particular user was posting negative comments across various subreddits, then it would be a strong indication that they are exhibiting trolling behavior. It is however important to keep in mind that all the subreddits on Reddit are regulated by moderators and it could be challenging for trolls to survive on these subreddits to begin with.

One of the goals of this study was also to evaluate how well a given subreddit promoted online learning and this can be answered using the code distribution. For our dataset, the code distribution shows that this subreddit ('askhistorians') promotes online learning with highest positive labels for code C3 (Explanation with Neutral Presentation), followed by C6 (Information Seeking), C7 (Providing Resources) and C5 (Socializing with Positive Intent). Similar inferences can be made about other subreddits as well based on the code classification of their posts. For future study, another recommendation is to assign a quality score for the subreddits based on the code distribution of that subreddit. For instance, if majority of the posts on a given subreddit are classified positive for codes with positive connotation, then that subreddit would have a higher score versus another subreddit where most of the posts are classified positive for codes that have a negative connotation. This would help the users evaluate the quality of posts on a particular subreddit as well as encourage moderators and participants to improve the score by providing valuable comments or replies to bring up the score. Users looking to learn from the posted answers would be able to get a good indication of the quality of responses in a given subreddit by just looking at the scores without having to read the comments/replies.

7. Appendix – A | Reddit Codebook

Code. Definition. Linguistic. Dialogue. Example

1. Explanation with Disagreement Expresses a NEGATIVE take on the content of the previous posts by adding new ideas or facts to discussion thread. ‘But’, ‘I disagree’, ‘not sure’, ‘not exactly’ with explanation/ judgment/ reasoning/ etc.

2. Explanation with Agreement Expresses a POSITIVE take on the content of the previous posts by adding new ideas or facts to discussion thread. ‘Indeed’, ‘also’, ‘I agree’, with explanation/ judgment/ reasoning/ etc.

3. Explanation with Neutral Presentation Expresses a NEUTRAL explanation/judgment/reasoning/etc. with neither negative nor positive reference to the content of the previous posts, nor necessarily any reference to previous posts. Posts with non-judgmental language. Advice, brainstorming and first-hand experiences are framed neutrally. ‘I can understand’, ‘interesting’, ‘depends on…’ or statement responses.

4. Socializing with Negative Intent Socializing that expresses negative affect through tone, words, insults, expletives intended as abusive. ‘no’, ‘you’re an idiot’, ‘this has been explained multiple times’

5. Socializing with Positive Intent Socializing that expresses positive affect tone, words, praise, humor, irony intended in a positive way. ‘thanks’, ‘great feedback’, ‘you’re correct’

6. Information Seeking Postings asking questions or soliciting opinions, resources, etc. (‘Does anyone know …?’ ‘How does this work?’). This does not include questions answered rhetorically within the post, e.g., if a question is asked and answered. ‘First you have to think what happens if …?’ and then you can see what happens’, ‘does anyone know’, ‘can anyone explain’

7. Providing Resources Postings that include direct reference to a URL, book, article, etc.; postings that call upon a well-known theory or the name of a well-known figure. Link to resource copied to post (book, URL, article, audio/video file). Referencing theory/theorists, scholar or public work (Einstein, Newton, Freud).

8. Subreddit Rules and NormsPostings on topics such as what is the appropriate sub-reddit for a particular discussion, what language is appropriate to use, how to back up claims by using resources, etc. ‘See/don’t forget subreddit link’, ‘this post doesn’t belong here’, upvote/downvote mentions, acknowledging OP redditors, and bots.

8. APPENDIX – B | SAMPLE FILE & GITHUB LINK

*T. **Sample File***

https://github.com/hirafatimaali/MRP/blob/master/askart.csv

https://d.docs.live.net/d312ffd7b7c8fce5/Documents/MRP/MRP/askart.csv

*U. **Github Link***

https://github.com/hirafatimaali/MRP

9. Appendix – C | List Of Fields In The Dataset

[1] "taskinfo__askhistorians_text"

[2] "taskinfo__askhistorians_score"

[3] "c1"

[4] "c2"

[5] "c3"

[6] "c4"

[7] "c5"

[8] "c6"

[9] "c7"

[10] "c8"

[11] "msg-id"

[12] "taskinfo__askhistorians_author"

[13] "taskinfo__askhistorians_comment_on"

[14] "taskinfo__askhistorians_date"

[15] "askhistorians_task_taskinfo__askhistorians_id"

[16] "taskinfo__askhistorians_parent"

[17] "taskinfo__askhistorians_parent_id"

[18] "taskinfo__askhistorians_submissions_author"

[19] "taskinfo__askhistorians_submissions_date"

[20] "taskinfo__askhistorians_submissions_downs"

[21] "taskinfo__askhistorians_submissions_id"

[22] "taskinfo__askhistorians_submissions_score"

[23] "taskinfo__askhistorians_submissions_text"

[24] "taskinfo__askhistorians_submissions_title"

[25] "taskinfo__askhistorians_submissions_type"

[26] "taskinfo__askhistorians_submissions_ups"

[27] "taskinfo__askhistorians_title"

[28] "taskinfo__askhistorians_type"

[29] "taskinfo__askhistorians_ups"

[30] "wordcount" – this is a calculated field

10. REFERENCES

[1] Tim Weninger, "An exploration of submissions and discussions in social news: mining collective intelligence of Reddit," *Social Network Analysis and Mining,* no. 10.1007/s13278-014-0173-9, 2014.

[2] H.-C. Kathy J. Liszka, "RedTweet: recommendation engine for reddit," *Journal of Intelligent Information Systems,* no. 10.1007/s10844-016-0410-y, 2016.

[3] S. A. S. S. Benjamin D. Horne, "Identifying the social signals that drive online discussions: A case study of Reddit communities," *Social and Information Networks,* no. arXiv:1705.02673 [cs.SI], 2017.

[4] P. Singer, E. Ferrara and F. Kooti, "Evidence of Online Performance Deterioration in User Sessions on Reddit," *PLOS ONE,* 2016.

[5] T. M. Rohlin, "Popularity prediction of Reddit texts," *San Jose State University, ProQuest Dissertations Publishing, 2016. 10128447.* , 2016.

[6] F. F. C. M. E. Z. M. S. Philipp Singer, "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?," *Social and Information Networks,* no. arXiv:1402.1386 [cs.SI].

[7] J. Paavola, T. Helo, H. J. Miika Sartonen and A.-M. Huhtinen, "The Automated Detection of Trolling Bots and Cyborgs and the Analysis of Their Impact in the Social Media," *European Conference on Cyber Warfare and Security; Reading,* pp. 237-244, 2016.

[8] "Discovering High-Quality Threaded Discussions in Online Forums," *Journal of Computer Science and Technology* , vol. 29, no. 3, pp. 519-531, 2014.

[9] J. Paavola and H. Jalonen, "An Approach to Detect and Analyze the Impact of Biased Information Sources in the Social Media," *European Conference on Cyber Warfare and Security,* pp. 213-219, 2015.

[10] T. S. a. D. Keim, "Visual Analysis of Social Media Data," *IEEE Computer Society,* no. 0018-9162/13/$31.00 © 2013 IEEE, 2013.

[11] M. P. Annie T Chen1, P. Shu-Hong Zhu2 and P. Mike Conway2, "What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques," *JMIR Publications,* no. J Med Internet Res 2015;17(9):e220, 2015.

[12] ,. M. T. Rudy Prabowo1, "Sentiment analysis: A combined approach," *Journal of Informetrics,* vol. 3, no. 2, pp. 143-157, 2009.

[13] A. K. U. Gunal, "The impact of preprocessing on text classification," *Information Processing and Management,* pp. 104-112, 2014.

[14] Z. S. S. A. K. M. Muhammad Usman, "Urdu Text Classification using Majority Voting," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 7, no. 8, 2016.

[15] "Analytics Vidhya," 26 Mar 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/. [Accessed 1 July 2017].

[16] "Wikipedia," This page was last edited on 7 May 2017, at 18:31.. [Online]. Available: https://en.wikipedia.org/wiki/Reddit. [Accessed 9 May 2017].