Answers to the Tasks

Aug 28, 2023

Group: Vec2R

Submission GitHub Repository: Click here

T1

Using PRAW, we scraped the top posts along with their comments on the Myanmar subreddit. These can be found in the 'posts.csv' file and 'comments' directory in the GitHub repository.

T2 and T3

The labels generated by the models and the majority labels are collected in the 'majority' directory in the GitHub repository.

T4

The rest of this document contains a summary of the Exploratory Data Analysis performed on the comments corpus and posts. The program for which can also be found in the GitHub repository.

T5 and T6

We randomly sampled 34 positive, 33 neutral and 33 negative comments from the comments corpus to perform human evaluation. Proper ethics were followed to annotate each comment by the three annotators.

T7

The Krippendorff coefficient (Krippendorff's alpha) = 0.6344.

The code for finding the coefficient is in the 'Krippendorff.py' file in the GitHub repository.

T8

The majority label given to a comment by the three annotators is also in the GitHub repository.

T9

There were a total of 5 comments that the model and annotators differed while labeling its sentiment. These comments are:

1) 'Lol, seriously dude?'
2) 'Here\'s a sneak peek of /r/AntiHateCommunities using the [top posts](https://np.reddit.com/r/AntiHateCommunities/top/?sort=top&t=all) of all time!\n\n\\#1: [Alt-right cesspool. Upvote so this is the first image that appears when you google "alt-right cesspool".](https://i.redd.it/zczx8vevckf61.jpg) | [98 comments](https://np.reddit.com/r/AntiHateCommunities/comments/lcw810/altright_cesspool_upvote_so_this_is_the_first/)
   #2: [Have you guys seen this sub? Seems like a hate sub parodying us.](https://i.redd.it/czzkcphfwoi51.jpg) | [116

comments](https://np.reddit.com/r/AntiHateCommunities/comments/iexxcy/have_you_guys_seen_this_sub_seems_like_a_hate_sub /)
#3: [We are the unseen heroes of the world](https://i.redd.it/b1u6ctq74j461.png) | [137 comments](https://np.reddit.com/r/AntiHateCommunities/comments/kaztkm/we_are_the_unseen_heroes_of_the_world/)(https://www.reddit.com/message/compose/?to=sneakpeekbot)
[Info](https://np.reddit.com/r/sneakpeekbot/)
[Opt-out](https://np.reddit.com/r/sneakpeekbot/comments/joo7mb/blacklist_viii/)

3) 'He probably will not shoot his workers like the military does, so the deal is still good.'
4) 'Ur mum m8 LULULULUL'
5) 'I know the OP through a mutual friend. He works at Tesla so I’m sure it’s legit.'

EDA on comments from Myanmar subreddit

August 28, 2023

Group: Vec2R

A. Dataset Description and Overview

We collected the comments from the top 100 posts from the Myanmar subreddit. We analyzed sentiment on the data we scraped from the subreddit and compiled the results here.

1) The total number of comments in the dataset: 3266
2) Sentiments detected:
    a) Positive: Comments through which the author is seen to be hopeful, appreciating and speaking on the positive aspects of the situation portrayed by the post.
    b) Negative: Comments that are abusive, express denial and negligence alongside hate is termed as a negative comment
    c) Neutral: Posts that aren't positive or negative.

In this document, we will present the results obtained from EDA and some observations.

B. Observations

1) Observations on the composition of the comments.

We analyzed the sentiments of various comments that were there in our dataset. While doing so, following observations were made:

a) Observations on length of comments:
Majority of the comments were less than 50 words long.

b) Observation on length of words used:
The mean word length of most of the comments was 7.

c) Most used stop words:
Figure 3 represents the frequencies of the top 10 stop words that were encountered in the data,

d) Most commonly used words that are not stop words:
The other most common words that were used are shown in figure 4. These results are also shown as a word cloud in figure 9.

e) Top bi and tri-grams:
Figure 5 and 6 represent the commonly occurring bi and tri grams.

2) Observations on the number of upvotes.

The Reddit API metadata labels the score for the posts and comments as the number of upvotes it received. We performed the following comparison of posts and comments based on their scores.

a) Posts vs Score:
From figure 7, we can infer that most of the posts had < 2000 upvotes.

b) Comments vs Scores:

From figure 8, we can infer that many comments had 0 upvotes. There were also some down votes. From figure 9, we are able to clearly tell that the majority of posts with a lot of downvotes (negative score) are mostly comments that were labeled as negative. Some neutral labeled comments also have negative scores. This in contrast to the positive labeled comments that have scores > 0.

C. Histograms and PDFs

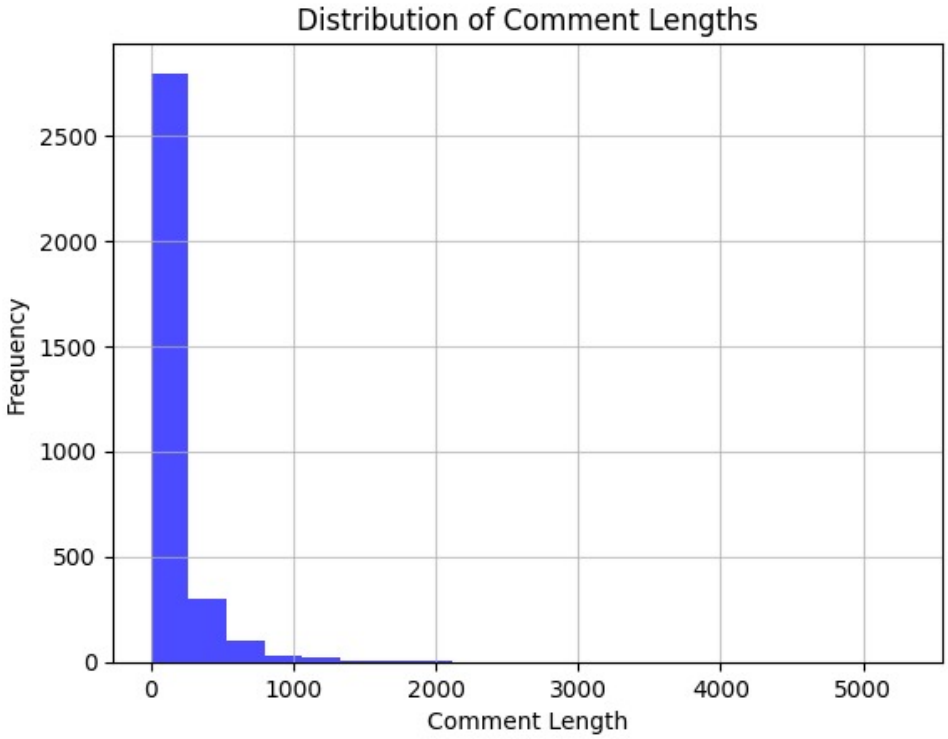

Figure 1: Frequency vs Length of comments

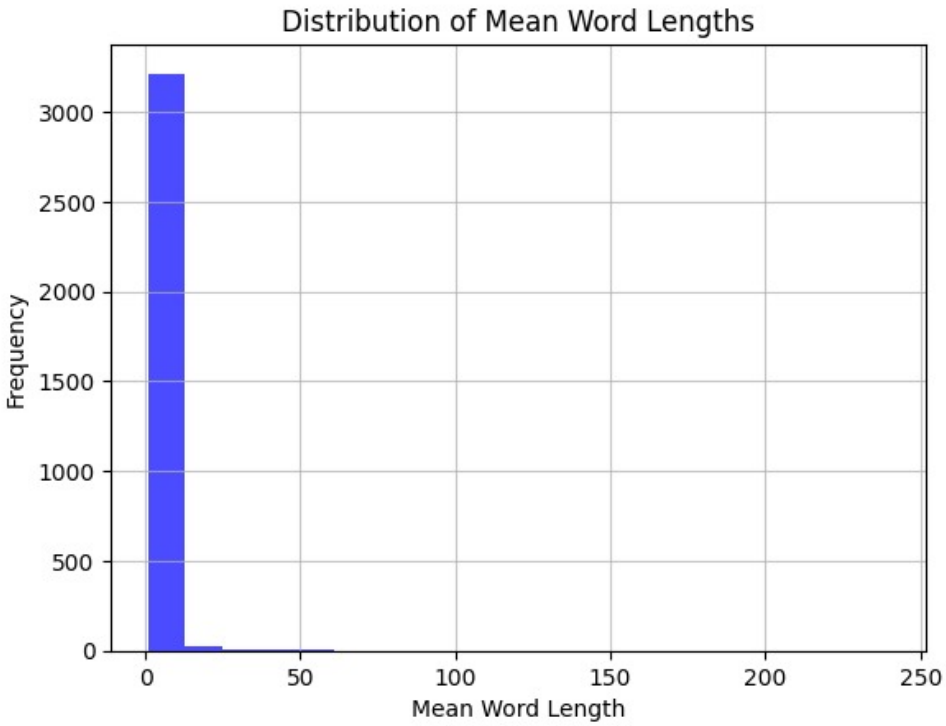

Figure 2: Frequency vs Mean word length

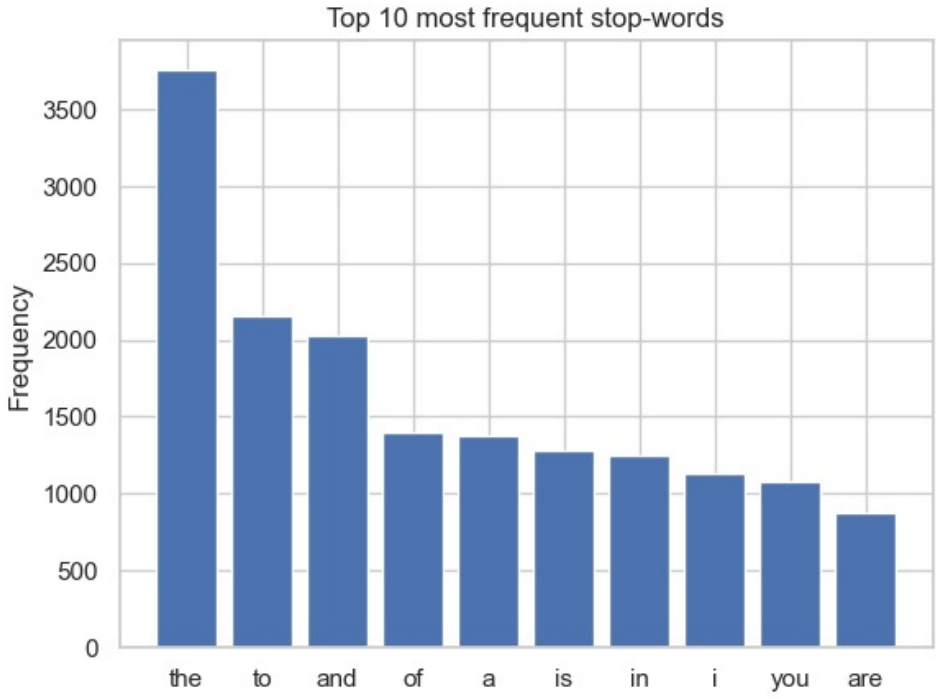

Figure 3: Frequencies of top 10 stop words.

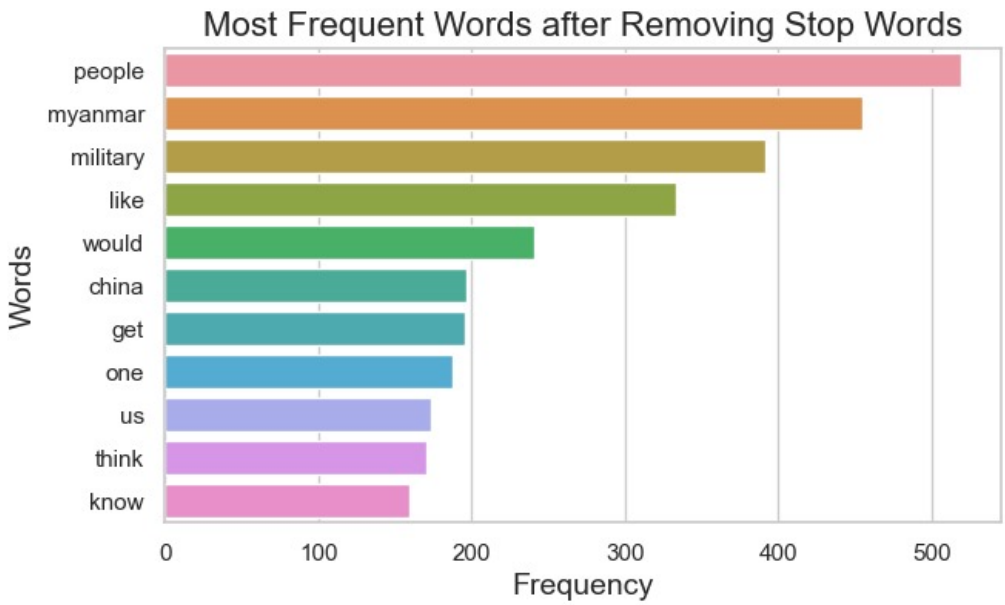

Figure 4: Frequency of most common words (after removing stop words from the corpus)

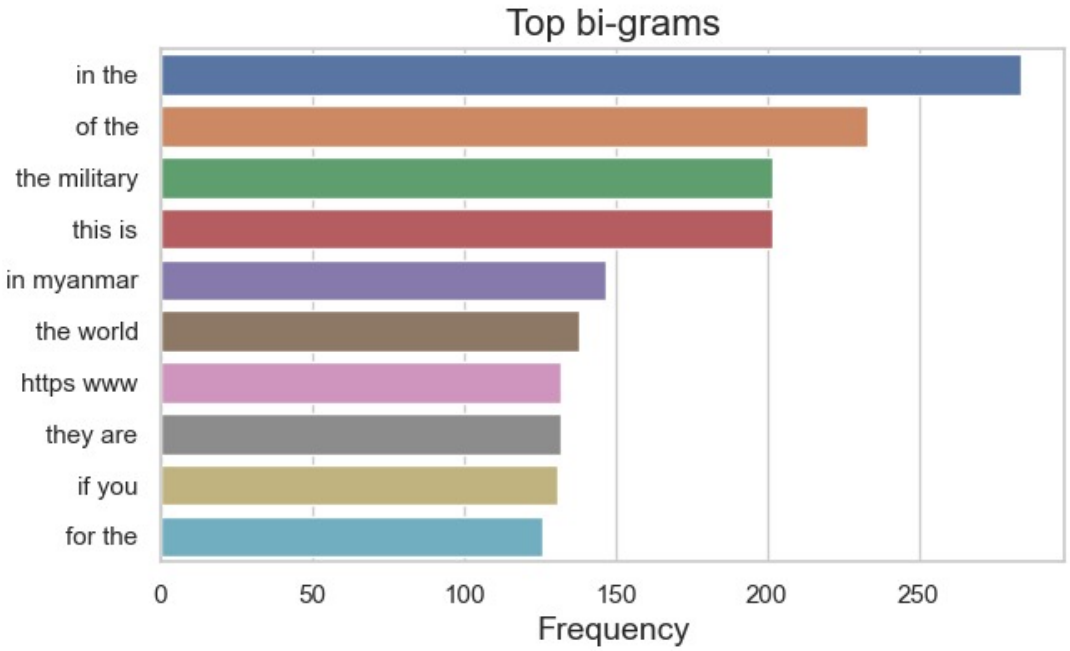

Figure 5: Top bi-grams in the corpus

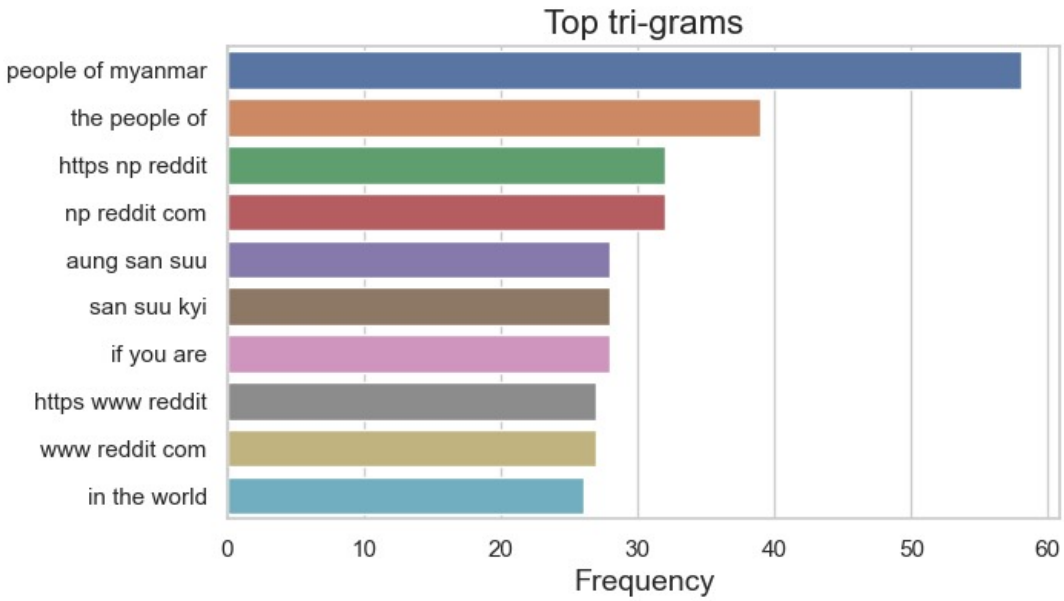

Figure 6: Top tri-grams in the corpus

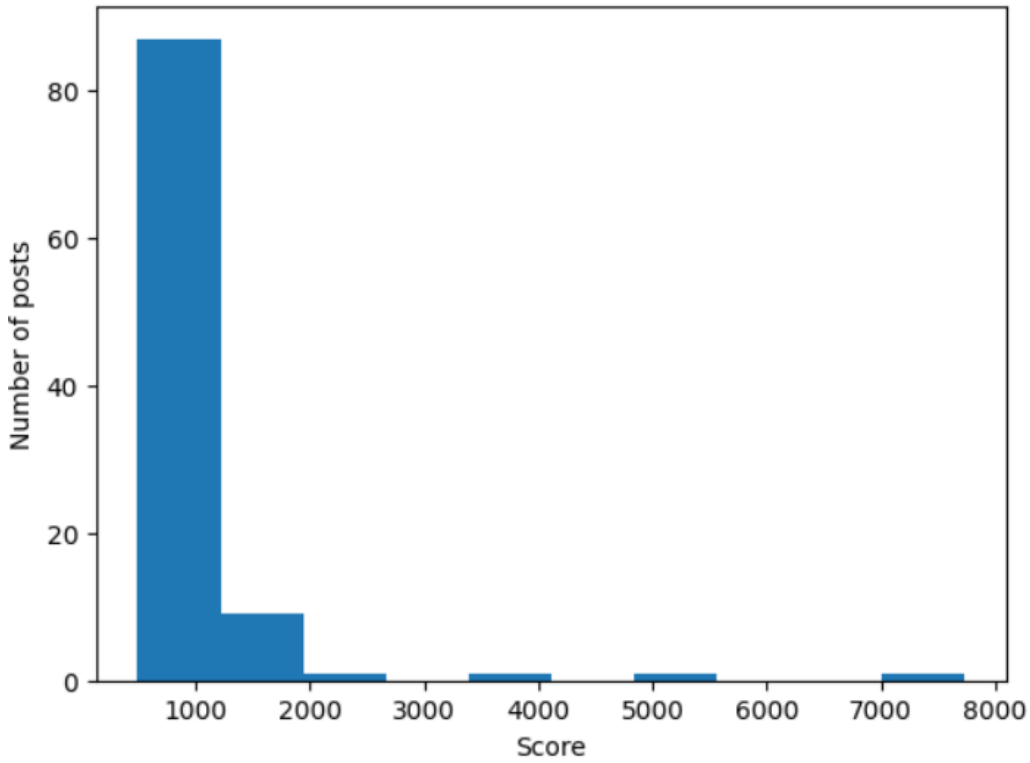

Figure 7: Histogram representing the number posts that received a particular number of upvotes.

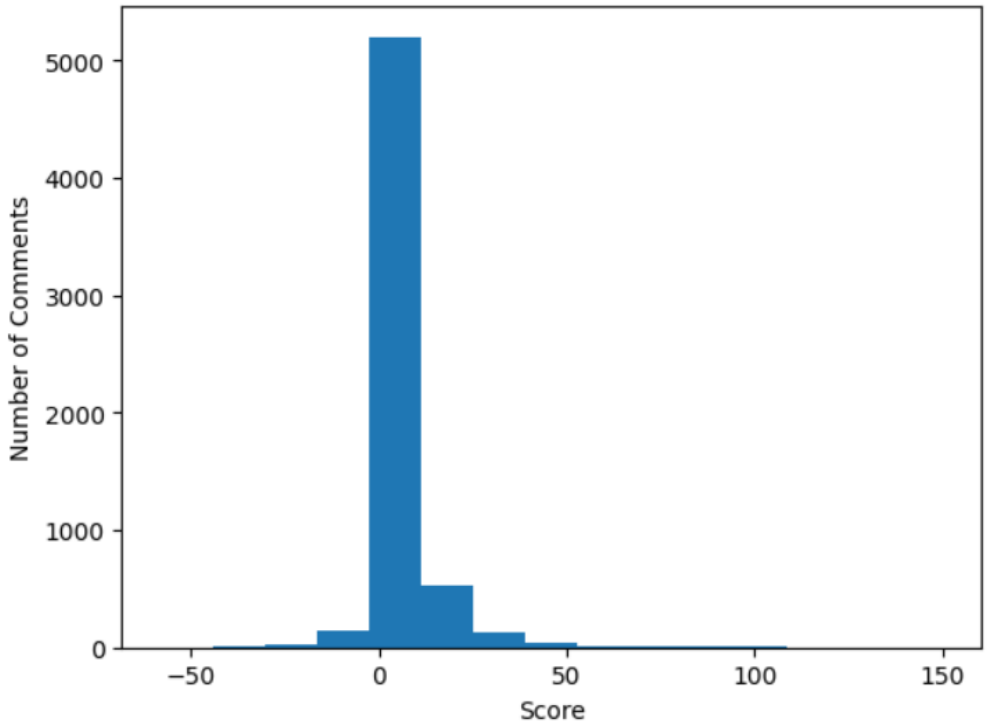

Figure 8: Histogram representing the number of comments that received a particular number of upvotes.

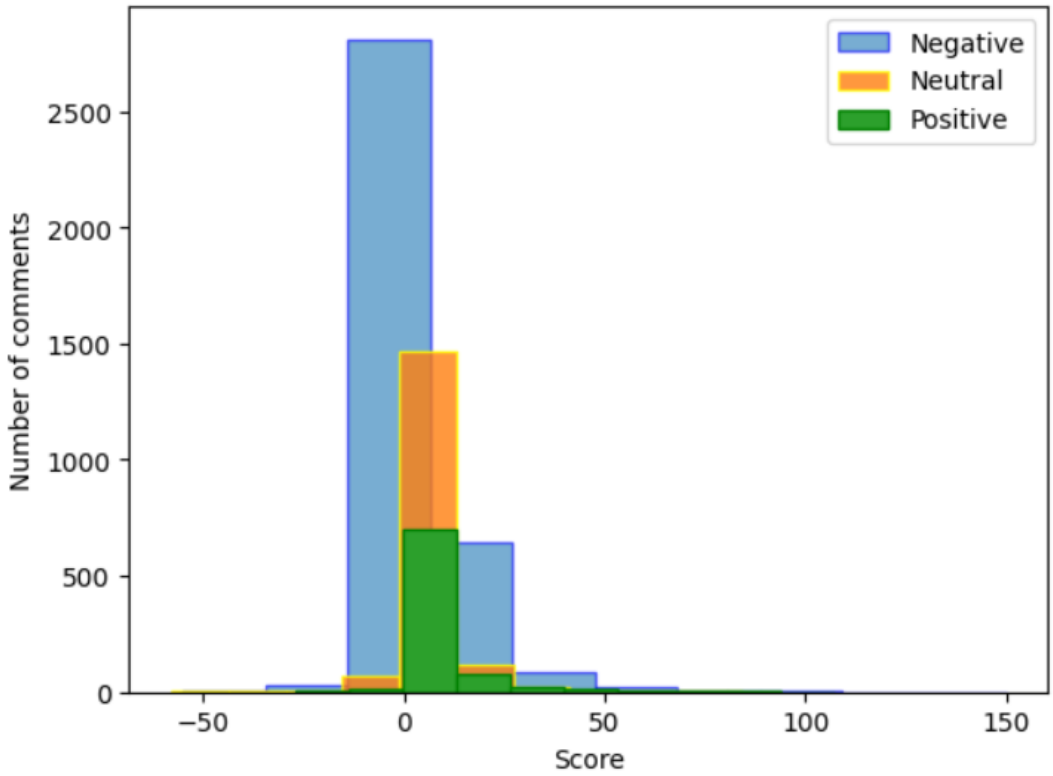

Figure 9: Histogram that represents number of comments vs score for all the three sentiments.

WordCloud:

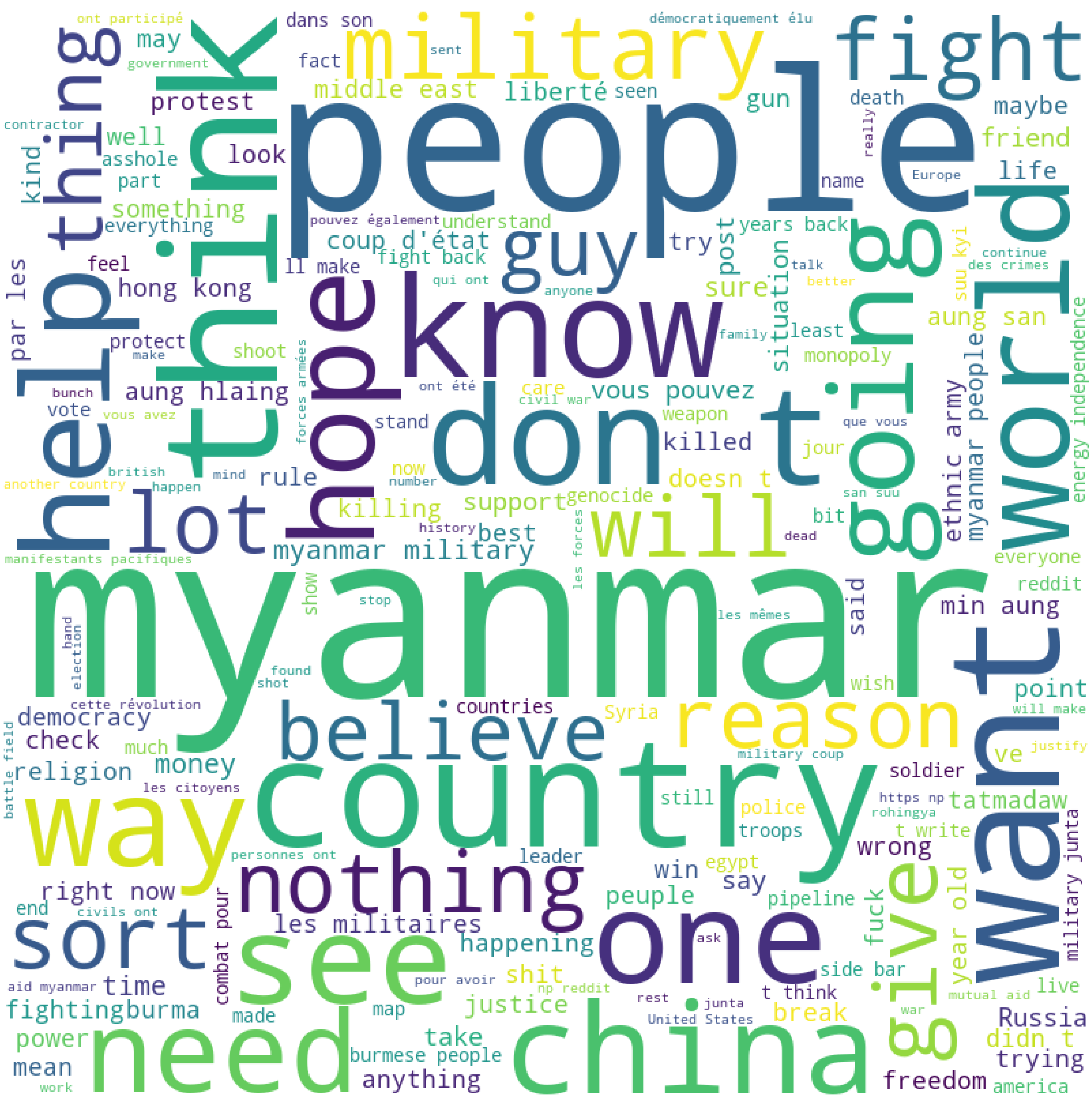

Figure 10: Word Cloud

D. References

1. https://colab.research.google.com/drive/1lno4DQA_bdT2Xd8cv3mwtRnwR1MdM-3u#scrollTo= w2frnq1OczG3
2. https://www.youtube.com/watch?v=Y7BSe7EiBTs
3. https://www.youtube.com/watch?v=FdjVoOf9HN4
4. https://praw.readthedocs.io/en/stable/
5. https://towardsdatascience.com/how-to-use-the-reddit-api-in-python-5e05ddfd1e5c
6. https://www.reddit.com/
7. https://www.datacamp.com/tutorial/wordcloud-python.
8. https://github.com/pskadasi/YT_Scrape
9. https://huggingface.co/models
10. https://github.com/pskadasi/eda_haberman
11. https://www.geeksforgeeks.org/generating-word-cloud-python/
12. https://neptune.ai/blog/exploratory-data-analysis-natural-language-processing-tools

Contributions

Name	Roll Number	Contribution (percentage)
Mithil Pechimuthu	21110129	11
Kaushal Kothiya	21110107	11
Dhruv Gupta	21110070	11
Rachit Verma	21110171	12
Sachin Jalan	21110183	11
Anish Karnik	21110098	11
Ayush Modi	21110039	11
Sahil Das	21110184	11
More Rutwik	21110133	11