Indian Institute of Technology Gandhinagar

Assignment 1

Data Scraping, Curation, and Analysis

CS 613: NLP

Group T10: Exquisiters

Aditi Dey (20110007)

Anavart Pandya (20110016)

Ayush Gupta (20110031)

Aditya Bhujbal (20110040)

Mann Jain (20110108)

Pinki Kumari(22210028)

Ronak Kalra (20110171)

Sakshi Jain (20110181)

Vedang Chavan (20110222)

Tasks:

1. Scrapping

Posts Metadata

Attribute	Description
`id`	ID of the submission.
`body`	Additional description text in post body
`date`	Date when the post was uploaded
`is_self`	Whether or not the submission is a selfpost (text-only).
`flair_text`	The flair's text content, or None if not flaired.
`num_comments`	The number of comments on the submission.
`nfsw`	Whether or not the submission has been marked as NSFW (over 18).
`score`	The number of upvotes for the submission.
`sticky_post`	Whether or not the submission is stickied.
`title`	The title of the submission.
`upvote_ratio`	The percentage of upvotes from all votes on the submission.
`url`	The URL the submission links to, or the permalink if a self post.

Comment Meta Data

Attribute	Description
`post_id`	ID of the parent post against which comment has been made
`comment`	The commented text
`upvotes`	Total number of upvotes/downvotes on the comment
`time`	Date time when the comment was submitted
`is_submitter`	Whether or not the comment has been posted by posts owner itself

2. Sentiment Analysis and Majority Label
    a. **Approach and Libraries**

        We used Pandas for data manipulation and the Transformers library from Hugging Face to access pre-trained sentiment analysis models.

    b. **Majority Voting**

        The code uses a majority voting system to assign a final sentiment label based on the output of three selected models. If all three models output different labels, the label given by twitter-roberta is chosen as the majority label.

    c. **Insights**
        i. Robustness and Reliability: The multi-model, majority voting approach enhances overall reliability, serves as an in-built quality control system, and is resilient against false positives and negatives.
        ii. Language Versatility: The model effectively labels sentiments in comments written in different languages, although it struggles with text in English that uses non-English vernacular (e.g., "Mujhe ye pasand nhi aya").

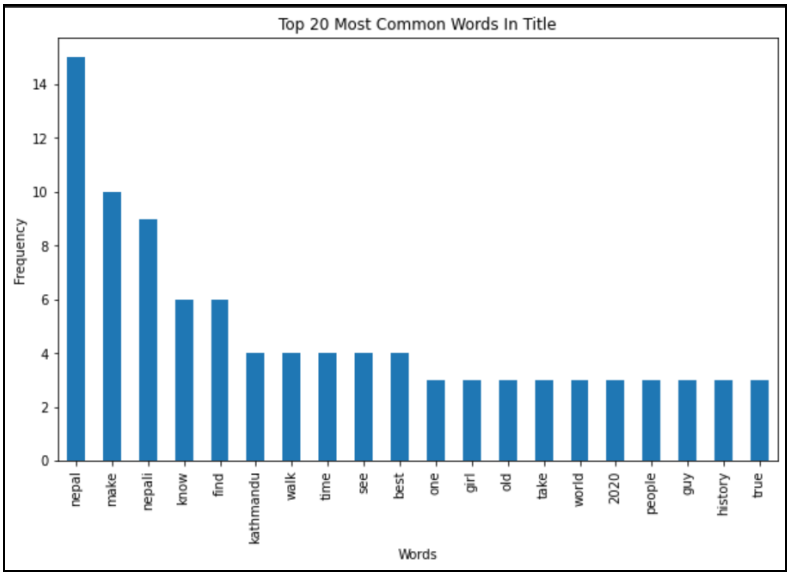
    d. **Considerations**
        i. Operational and Contextual Constraints: The method requires higher computational resources, lacks a tie-resolution mechanism, and may miss nuanced meanings in specific linguistic or cultural contexts.

3. Majority label for the entire corpus

    The majority label corresponding to each comment has been calculated using pandas dataframe mode function. The majority label for the entire corpus is neutral.

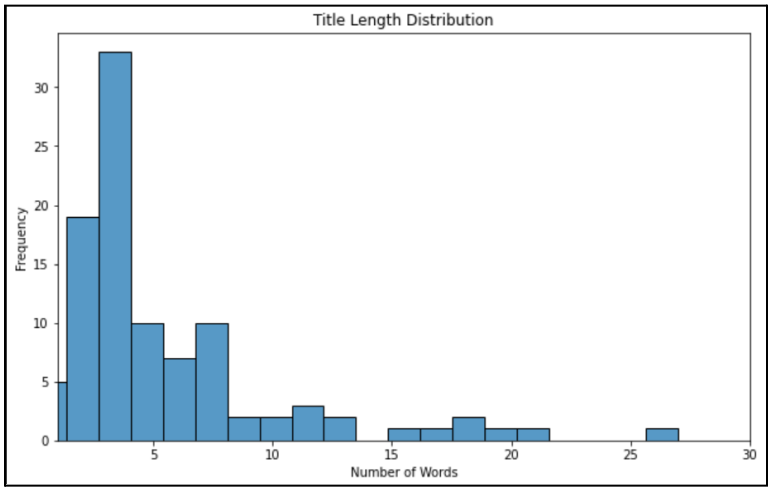
4. Preprocessing and EDA

- Preprocessing steps include removing urls and special characters, not word characters, spaces or periods, converting to lowercase and removing extra spaces between two words.
- The number of unique words in the entire corpus is 13100, and mostly the words are of length 3-10.
- **Frequency of Words in the corpus**



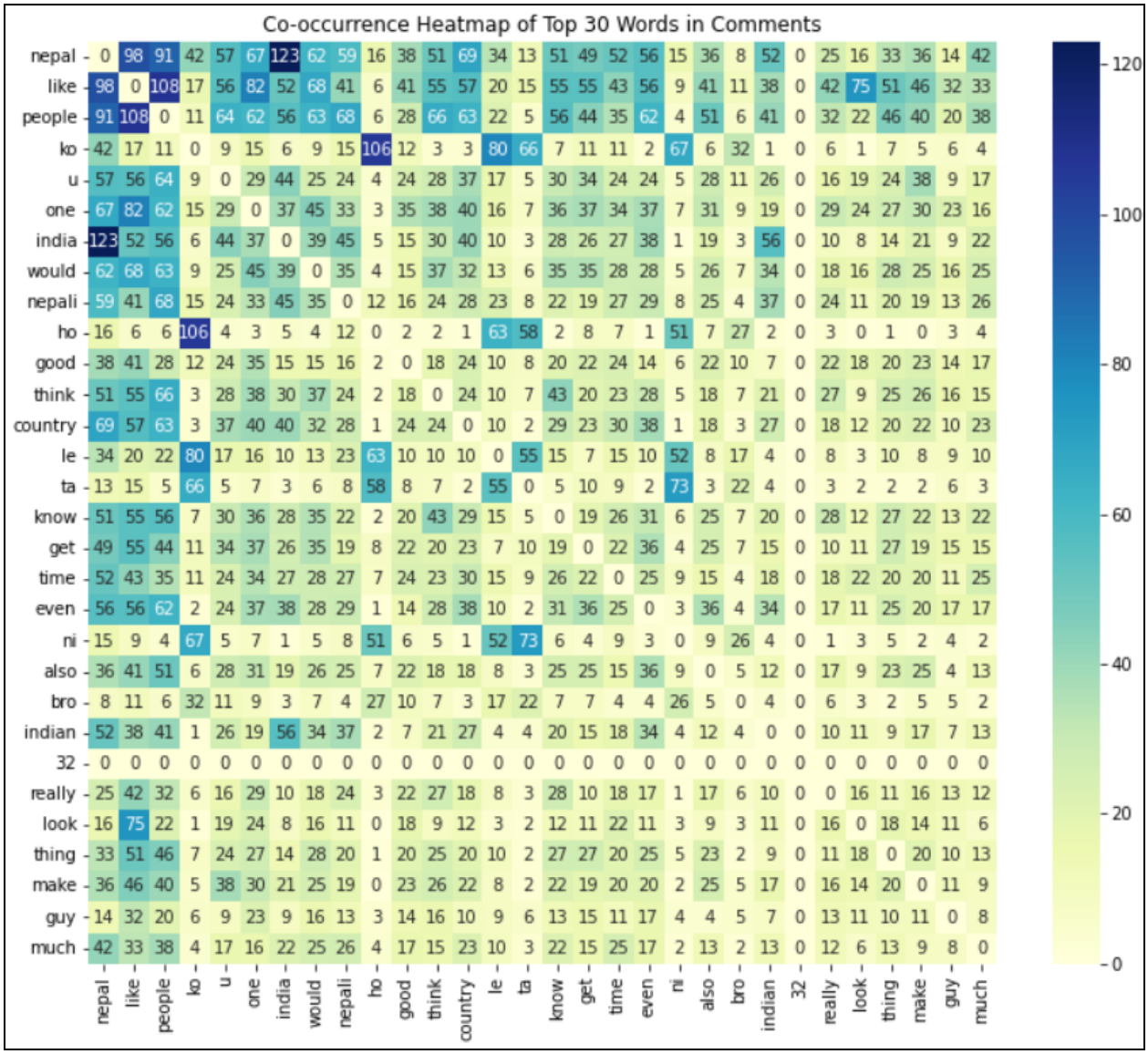
We took the top 20 words from the posts_segmented and we plotted their corresponding frequencies as shown in the graph. We see that 'nepal', 'make' and 'nepali' are the most frequently occuring words in the corpus, as we can see in the Word Cloud.

- **Length of titles**



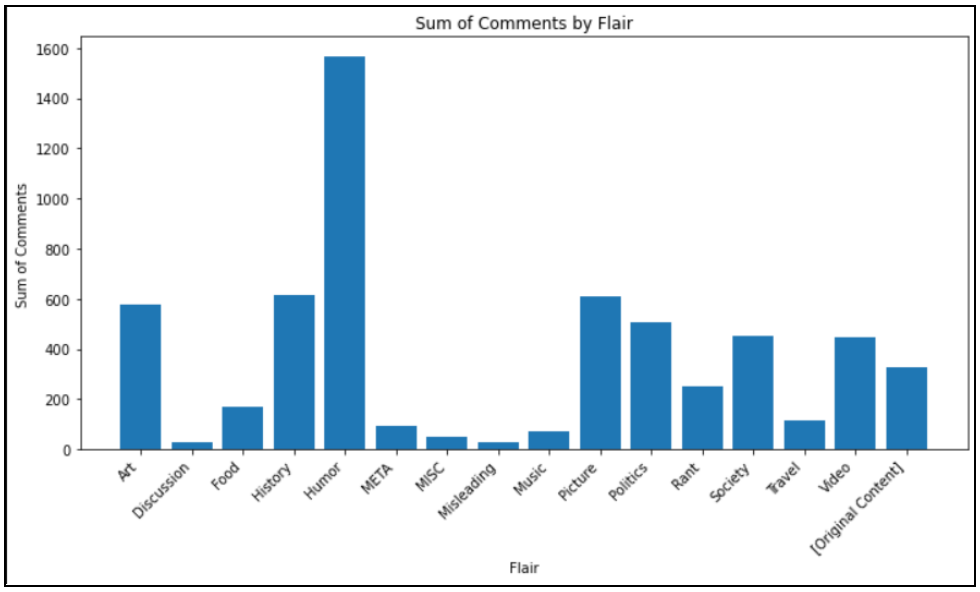
We analyzed the length of the titles of all the posts. and observed that most of the titles had lengths up to 5 as long words are rare.

- **Co-occurrence matrix**



In the heatmap, we notice that the diagonal elements are predominantly zero. This suggests that identical words, such as "Nepal Nepal" or "India India," rarely appear together within the corpus. When two words show a more pronounced value at their intersection in the co-occurrence matrix, it signifies a higher probability of them being used together.

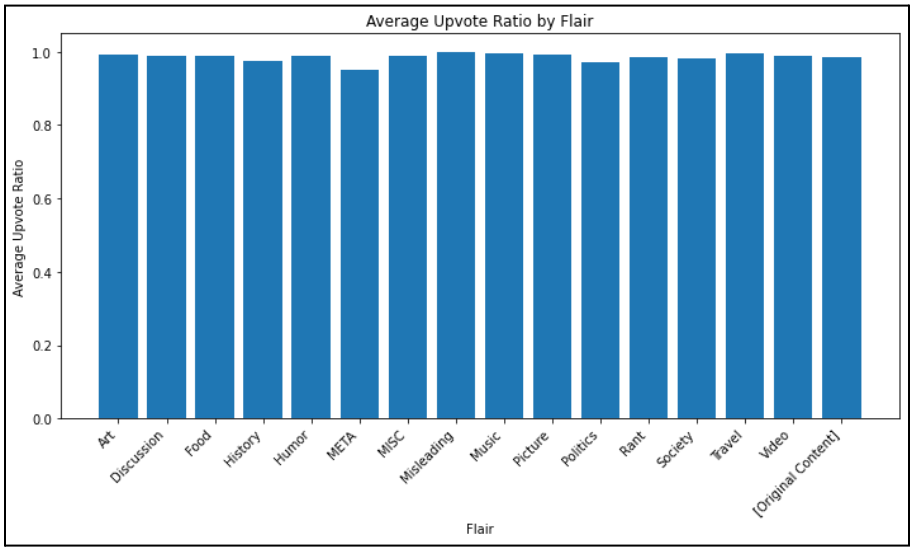
- **Sum of Comments by Flair**



"Sum of Comments by Flair" refers to the cumulative count of comments contributed by users who possess a specific distinctive "flair" associated with their usernames. This metric can be utilized to monitor and assess the level of engagement or activity exhibited by users with particular expertise, interests, or roles within an online community. It offers insights into the prominence and impact of users participating in discussions related to specific subjects.

Observations:

1. Predominantly, the flairs are categorized as "Humor," followed by "Misleading," "Travel," "Music," and others.



The average upvote ratio, calculated as the total upvote ratio divided by the total posts for each flair, demonstrates remarkable consistency across all flairs. This suggests that each flair garners a similar level of audience engagement and reactions.

4. Sampling

   The sampling is performed by grouping the comments in Positive, Negative and Neutral sentiments and then randomly choosing an equal number of comments among the three groups.

5. Human Evaluation

   Three annotators labeled the 100 randomly sampled comments into 'positive', 'negative' and 'neutral' based on their individual understanding of sentiments.

6. Krippendorff's alpha

   Krippendorff's alpha ranges between -1 and 1. Here, -1 represents agreement between the annotators is less than expected by chance whereas 1 represents perfect agreement. We have used 'Quica' to calculate the Krippendorff's alpha score for amongst all annotators and pairwise combination of annotators which are as follows:

   Overall Agreement Score: 0.54544

   Agreement Score b/w annotator-1,2: 0.54105

   Agreement Score b/w annotator-2,3: 0.52703

   Agreement Score b/w annotator-1,3: 0.57029

7. Displaying 5 Comments

Comment Id	Comment	Model Majority label	Human Majority label
454	He is a bold one	neutral	positive
481	Per kera Rs.500. Another bonus.	positive	neutral
564	This needs more upvotes.	neutral	positive
764	But those two guys didn't break any law.	negative	positive

928	There are 67 ethnic groups.. Haven't you studied social studies? And this logo doesn't even represent many high percentage population cultures.. like Tamangs.. Rais.. Maithali.. Magar.. Gurung.. etc etc Also this logo will be good for subreddits like r/Khas or something.. but if we put this as a logo of r/Nepal then it'll create division among ourselves.. And this is against our own country.. i.e. gardens of all types of people.. Edit: sorry it's not 67.. but 123 ethnic groups.. my bad for not researching properly..	negative	neutral

Results:

- The number of unique words in the entire corpus is 13100, and most of the words are of length 3-8.
- We took the top 20 words from the posts_segmented and we plotted their corresponding frequencies. We see that ‘nepal’, ‘make’ and ‘nepali’ are the most frequently occuring words in the corpus, as we can see in the word cloud.
- **Co-occurrence matrix**
    - We observe that the diagonal elements are mostly zero in this heatmap, so we can say no two same words appear together in the corpus. Example - nepal nepal, india india, etc.
    - If two words exhibit a higher value at their intersection within the co-occurrence matrix, it indicates a stronger likelihood of them occurring together.
- **Sum of Comments by Flair**
    - Mostly flair is of Humor type, and thereafter art, history, picture, politics, society, etc.
    - The average upvote ratio [total upvote ratio / total posts for that flair] for all the flairs was nearly the same, i.e., each flair has the same amount of audience reach and reactions.

- **Human Evaluation w.r.t Sentiment Analysis**
  - The first two models(Twitter-Roberta and Bertweet Base) struggle with grasping subtle linguistic cues or sarcasm, leading to discrepancies.
  - Humans recognize the indirect positive sentiment related to the desire for approval and popularity which are not be captured by models in some cases.
  - Emphasis placed on specific words through punctuation or capitalization alter sentiment which seems to be not recognized by first two models.

Word Cloud of the entire corpus of comments after removing the stop words and considering the minimum length of words as 3.

Work Distribution Percentage:

- **Aditi Dey : 11%**
- **Anavart Pandya : 11%**
- **Ayush Gupta : 11%**
- **Aditya Bhujbal : 11%**
- **Mann Jain : 11%**
- **Pinki Kumari : 11%**
- **Ronak Kalra : 11%**
- **Sakshi Jain : 11%**
- **Vedang Chavan : 12%**