CS 613: NLP

Assignment 1: Data Scraping, Curation, and Analysis

Total marks: **100 Pts.** **(+10 BONUS)**	**Submission deadline: 23:59:59 Hrs, August 28, 2023**

Assignment Instructions

A walkthrough of the assignment will be presented in the upcoming Monday's class on 21 August 2023:

1. Regarding the late submission, we will be following the penalty as per the table:

Late Submission	Penalty (Out of 100)
Till 1-hour past the deadline	5 points
1 to 12 hours past the deadline	10 points
12 to 24 hours past the deadline	20 points
24 to 36 hours past the deadline	40 points
36+ hours past the deadline	100 points

2. We will follow the zero plagiarism policy, and any act of plagiarism will result in a zero score for the assignment.
3. Please cite and mention others' work and give credit wherever possible.
4. If you seek help and discuss it with the stakeholders or individuals, please ask their permission to mention it in the report/submission.

Problem Statement

Analyze the sentiments of comments for top posts of all time in India and its neighboring countries' subreddits.

Tasks (80 Pts. +10 Pts. [BONUS])

1. Scrape comments from the top 100 posts (of all time) of the below subreddits. Pick Subreddit as per your Team Id. here (T1 will choose Afghanistan, T2 will choose China, and so on.) **[25 Pts.]**
    a. Subreddits
        i. Afghanistan: r/afghanistan
        ii. China: r/china
        iii. Bhutan: r/bhutan
        iv. Bangladesh: r/bangladesh
        v. India: r/india
        vi. Sri Lanka: r/srilanka
        vii. Pakistan: r/pakistan
        viii. Maldives: r/maldives
        ix. Myanmar: r/myanmar
        x. Nepal: r/nepal
    b. Refer: https://github.com/pskadasi/Reddit_Scrape
    c. Note: Using PRAW, there is a feature to scrape the top posts of the subreddit of all time.
2. Pick the 3 most popular models for sentiment analysis from Hugging Face and run inference on these comments extracted [NO NEED TO TRAIN THE MODEL] to generate labels (positive, negative, neutral) for the entire corpus. **[5 Pts.]**
    a. Models: Below are the sentiment analysis models which have the highest #downloads on Hugging Face
        i. cardiffnlp/twitter-roberta-base-sentiment-latest
        ii. finiteautomata/bertweet-base-sentiment-analysis
        iii. Seethal/sentiment_analysis_generic_dataset
3. Get the majority label by performing the majority vote on these 3 labels corresponding to every sentence to get a single label for the entire corpus. **[2 Pts.]**
4. Preprocess the data and do Exploratory Data Analysis (EDA) on these comments for the entire corpus. You can also use other useful parameters in the extracted corpus file like #upvotes for the comment, and others, etc. (explore subreddit API object, check references for other metadata and GitHub) **[5 Pts.]**
    a. Get an idea of how to do EDA from here: https://github.com/pskadasi/eda_haberman or you can refer to any Kaggle notebooks.
5. Now, randomly sample 100 sentences (annotated comments with majority label) for the entire corpus. Make sure there is an equal proportion of sentiments (pos, neg, neu), the dataset should be balanced. **[3 Pts.]**
6. Do human evaluation on these 100 sentences i.e., annotate these 100 sentences for sentiment by 3 different annotators from the team. **(Please be very careful with the annotations, annotations should be done in isolation with one another. After annotation is done, one shouldn't change**

their annotation after one sees others' annotations for a particular sentence. Proper ethics should be followed as human evaluation is very very important in NLP research). **[35 Pts.]**

7. Find the inter-annotator agreement using Krippendorff's alpha for the sample of 100 sentences. (it's completely fine even if the agreement is low, but report genuine agreement value in the PDF file.) [BONUS] **[+10 Pts.]**
8. Perform a majority vote of the three annotators' labels to get the majority label. **[2 Pts.]**
9. Display 5 comments from the sample of 100 sentences where the majority label of models and majority label of human annotations is different, if present. **[3 Pts.]**
    a. Show with proof (numbers/label distribution etc.) and answer, if not present.

Points Split: 25+5+2+5+3+35+(10)+2+3 = 80 + 10 (Bonus)

Results (20 Pts.)

As per the analysis done, answer the following question in a document:

1. Write your analysis based on EDA and human evaluation with respect to sentiment analysis. (100W) **[10 Pts.]**
2. Create the Word Cloud of the entire corpus using the Word Cloud package. **[10 Pts.]**
    a. Remove the stop words.
    b. Set min_word_length as 3

Submission

1. Submit a CSV file for posts.
2. Submit a CSV file for comments, its model predicted labels, majority label, metadata - Number of CSV files for comments are 100, check sample submission folder in point 7.
3. Submit a CSV file of human-evaluated 100 sentences with individual annotations by 3 annotators.
    a. The column name should be named after the student-id of the respective annotator.
4. A PDF file of answers for the Tasks and Results section.
5. Write the individual contribution percentage in the document.
6. An iPyNB file or a link to the colab file where you have done EDA, the rest of the analysis of the dataset with proper documentation in the Python notebook (you can add this link in the PDF document itself)
7. A sample submission file: CLICK ME
8. Submit the ZIP file of the above submission folder: CLICK ME

Expectations from the team:

1. Properly divide the team into sub-groups and distribute your tasks equally.

References

If required, please feel free to take help from the following 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

Reddit API Metadata:

Author Information:

Username: The username of the user who posted the content.
Author Karma: The total karma score of the user.
Account Age: How old the user's account is.

Submission Information (for Posts):
Title: The title of the post.
URL: The URL linked in the post.
Score (Upvotes): The number of upvotes the post has received.
Number of Comments: The total number of comments on the post.
Created Timestamp: The time the post was created.
Subreddit: The subreddit where the post was made.

Comment Information:
Comment Text: The content of the comment.
Comment Score (Upvotes): The number of upvotes the comment has received.
Parent Comment ID: The ID of the parent comment (for nested comments).
Comment Depth: The depth of the comment in the thread.
Created Timestamp: The time the comment was created.

User Engagement Metrics:
Upvote Ratio: The ratio of upvotes to the total votes (upvotes + downvotes) on a post.
Award Count: The number of awards (e.g., Reddit Gold, Silver, Platinum) a post or comment has been received.

Media Information:
Media Type: Whether the content includes images, videos, links, etc.
Media URL: The URL of the media content.

Text Length:
Text Length of Post: The number of characters in the post's title and body.
Text Length of Comment: The number of characters in the comment's text.

Contextual Information:
Post ID: The unique identifier of the post.
Comment ID: The unique identifier of the comment.
Parent Post ID: The ID of the post to which the comment is responding.
Subreddit Category: The category or topic of the subreddit.

Flair Information:
User Flair: Custom labels associated with a user.
Post Flair: Labels assigned to posts to categorize or tag them.

Reddit Awards:
Award Type: The type of award given to a post or comment.
Award Count: The number of times an award was given.

Link Information:
URL Shortener: Whether the URL is a short link (e.g., bit.ly).