Made By Prabhsimran and Vidham

EDA for our Russia Ukraine war Dataset

1 Column_id -> unique values
This is unique id for our dataset.

2 Score -> It is total upvotes - downvotes. We will plot a visualization predicting the frequency of each score.

3. Self_text -> It consists of all the ~~texts~~ comments in our dataset. In EDA, we will try to explore the comments & try to find what all needs to be done in data cleaning. Once done, we will compare cleaned comments with the original ones to check if data cleaning is done correctly or are there ~~any~~ is unnecessary cleaning happening.

4. Subreddits -> We will check the number of all the subreddits. Then we will count the frequency to see the most popular subreddit.

5. Created_time -> It tells us when comments was created. Though, this column is not relevant now because our current model is not able to use context (time of comment, real world news at that time). This column is not much used in current situation but can be useful to analyse

when most of the ~~active~~ users are active.

6 Post-id → It contains id of post under which comments are being posted. There are 55000 unique values. Also we can plot the frequency of each post id column to get which topic is being the most discussed under each post

7. author-name → It contains name of the person who wrote the comment. It seems ~~of~~ to be not much use; but using this column and the already present list of bot users that can be found on the internet, we can eliminate the bots from the dataset so as to strengthen the integrity of our data and analysis

8. Controversiality - It consists of 0 & 1.
0 → non controversial
1 → controversial
By using count function on this column we can check whether our data is imbalanced or not because according to our situation of controversial or not, no more than 5-10 % of the posts should be controversial

9. Ups → No. of upvotes received by a comment

10. Downs → No. of downvotes received by a comment.

classmate
Date
Page

9. We will be predicting controversiality of new posts, so it won't be carry any upvotes and downvotes. Therefore, we are still not sure, whether to use these columns for our feature engineering or not.

10. User_is_verified → In this Column we will analyse all the true values and delete all the rows that have false user verified values because if we want to maintain simplicity our data.

11. Users_account_created_time → Using year and month of this column, we will plot the years of the account created we will have time series chart depicting the frequency of the account created. It will helps us to have a glimpse on the type of audience that is most active on reddit.

12. User_awardee_karma, user_awarder_karma, user_link_karma, user_comment_karma, user_total_karma

These 5 columns are mostly same, so instead of using each column we will go with the user_total_karma so as to not ~~of~~ overfit the model. Then we will perform count function on user_total_karma to understand the karma distribution of the users. We have to use groupby on user_id to get user karma (We need to explore whether user_total_karma doesnot "gives"* skewed result and which karma column is most suited. As of now * we will go with total karma., Later we will try diff. combinations of karma.

13) Post_score -> It is sum of likes ~~and dislikes~~ of all the comments under the post. Using freq destribution for each post id we can plot graph of total score of each post

14) post_self text -> It contains the captions made by the post creater for his own post. ~~From EDA~~, we found that ≃86% of this column is empty, so we have to consider if this column is relevant for our task or not

15) post_title -> It is the text of the post associated with the post id. It contains the title text under which the comments were made using length function on this column and the length of each ~~post~~ text for each title then we plot ~~freq. of all posts~~ graph to compare length of each title text

16. Post_upvotes_ratio -> Using this column, we can find the most popular comment among all the comments made under the Same post.

17. Post_thumbs-up → Same as post_score.

18. post_total_awards_received -> It is all zeros. Therefore, not a relevant column.

19. post_created_time -> This column ~~will~~ is ~~be~~ used to know when each post was created and also when the most popular posts were created.