DATA PROCESSING AND ANALYSIS

1. DATASET:

    1. Hashtags crawled: #everydaysexism, #genderbias, #genderstereotype, #heforshe, #mencallmethings, #metoo, #misogynist, #notallmen, #questionsformen, #slutgate, #wagegap, #weareequal, #womenareinferior, #workplaceharassment, #yesallwomen
    [Stored in hashtag_name.csv files]

    2. Total no. of posts crawled : 12484 (10 comments per post)
    [Stored in Merged_Data.csv]

2. DATA PRE-PROCESSING:

    1. Tokenised text using regex tokeniser.
        - Regex used: [a-zA-Z0-9_#]+
        - Removes text written in languages that do not use English alphabets like Hindi, Urdu, Arabic, Chinese, Japanese, etc.
        - Removes punctuations other than _, # (Kept hashtags as it is)
        - Removes emojis
    2. Removed numbers from the text
    3. Stopword removal
    4. Lemmatisation
    5. Case folding

    [Stored in Processed_Data.csv]

3. DATA FILTERING:

    1. Data can be filtered using one of the following criteria:
        1. Restrict length of text in post (text = hashtags + caption + comments)
        2. Separate restrictions on number of hashtags in post and length of other text in a post. (other text = caption + comments)
        3. Separate restrictions on number of hashtags in post, length of caption in a post and length of comments in a post.
    2. Used 3rd criteria and filter data by putting restriction on minimum number of hashtags in a post, minimum number of words in caption and minimum number of words in comment.
    3. Analysis of filtered data size is done where threshold values are varied as follows:
        1. Minimum no. of hashtags : 1-10
        2. Minimum no. of words in caption : 5-10
        3. Minimum no. of words in comments : 5-10

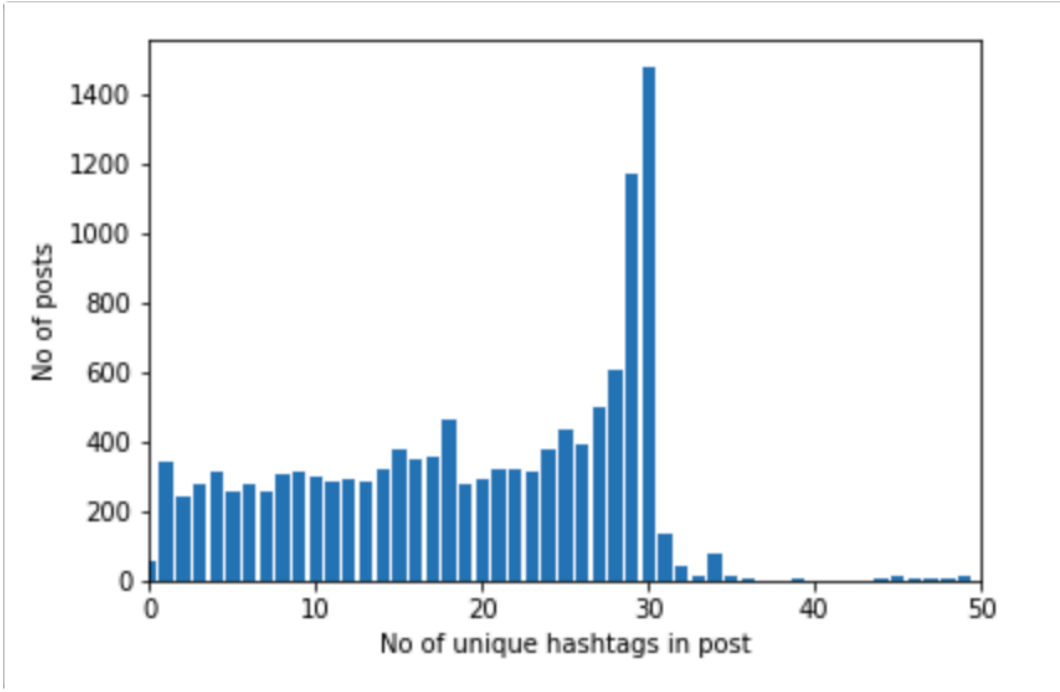
        [Stored in Filtered_Data_Analysis.csv]
    4. From above analysis, it is observed that filtered data with parameters (min no. of hashtags = 10, min no. of words in caption = 10 and min no. of words in comments = 10) have approximately 3000 posts which is 1/4th of the original data.
    [Stored in Filtered_Data.csv]

4. DATA ANALYSIS:

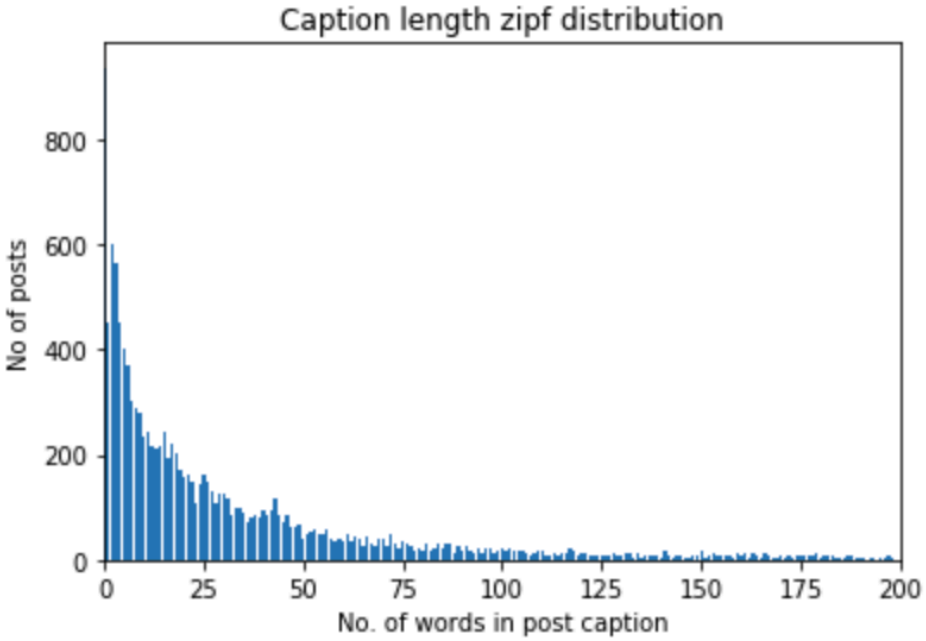
- Mean and median of caption length, comments length and total text length (both caption and comments) is calculated.

- Below graphs will be useful to decide range of threshold values for filtering data as well as to gain insights regarding how much trimming of text will be required if model can accept only fixed length input.

    1. No. of unique hashtags in post text content VS No. of posts in corpus containing 'x' No. of unique hashtag



2. No. of words in post caption VS No. of posts in corpus containing 'x' No. of words in caption



3. No. of words in post comments VS No. of posts in corpus containing 'x' No. of words in 10 comments

---