Final Project Individual Report

Shuting Cai

Dataset

The data can be extracted using Reddit API, sourced from Pushshift. Since the Pushshift API limits the number of 1000 data per request, it is necessary to apply multi-threaded downloads. PMAW (Pushshift Multithread API Wrapper) is a wrapper for the Pushshift API which uses multithreading to retrieve Reddit comments and submissions.

There are nearly 100k observations from January 2021 through the end of October 2022. Some data preprocessing methods are applied to the dataset for EDA, including changing all words to lowercase, tokenization, removing stop words, and lemmatization using the NLTK package.

Mywork

Digging into the data, the percentage of data showing real incidents of MacOS and Windows in the submission is 0.50 while the ratio of data showing real incidents MacOS and Windows is 0.55:0.45. Taking a quick look at the histogram plot, the observation of these two subreddits are half and half perfectly at the time period we taken the data, it indicates that the user activity and the amount of user in these two operating systems are very similar. That's the reason why the team would like to take this two subreddits.

I did EDA part to us know better the context of the submission and comments for these two subreddits. The histogram plot of the subreddit in submission and commetns, with a basic idea about the submission and comments before starting sentiment analysis. it is indicated that the sentence length and word count in the submission is higher than the comments. Most often, the submission person, known as a blogger needs more sentences and words to describe a thing he/she wants to talk about and share ideas, the viewer uses one word or one sentence to leave comments. The average word for each sentence is not a difference between submission and comments. In general, a normal word length is 4.7 characters,

I was checking the most common words for these two subreddits in submission and comments, showing the results with bar plot and also made one fun word cloud to show the most common words in two subreddits.

Checking with the most common word in subreddits, it can help people know the characteristics of Windows and MacOS and help Microsoft and Apple understand the demand and feedback of the user in these two operating systems. There are a couple of keywords in the bar plot and the word cloud that "please," "help, " "issue," "support" from windows users and "apple, " "time," "try," and "weird" that conclude there are a lots of windows users would like to talk about the problem they meet and seeking solutions to solve it.

Before getting to the model, I tried to use Textblob to do sentiment analysis, which is using a bag of words classifier. This method can get the sentiments of the text without modeling, and it will give the sentiments and also the polarity of the sentiments. It would help us to have a fundamental idea on the further modeling building. Exploring in HuggingFace and doing some research to find the models about Roberta base model trained on about 58m Tweets from Twitter and Distilling BERT base trained on IMDB dataset.