Topic Proposal

Team 2 - Yue Li, Aihan Liu, Shuting Cai

Reddit is a platform for users to spontaneously discover and share content while other audiences can vote or comment to show their opinions on the submitted post. Categories of content on the site are known as "subreddits," which is like a community organized for those people who have similar interests, content on the subreddit includes news, video games, movies, music, books, fitness, food, and image sharing, among others (https://nealschaffer.com/subreddit/).

In our project, we are going to take submission posts and comments from two subreddits of Apple and Windows to classify the related keywords difference between them, to predict the predominant sentiment comments on the submitted post with the pre-trained model of Twitter Roberta base sentiment and distilling BERT base trained on IMDB data. With the application of integrated gradient, the importance of the word or sequence on the predicted sentiments would be determined. The class of emotion would be predicted as positive, negative, and neutral.

Dataset Description

The Reddit data can be extracted using Reddit API, sourced from Pushshift. There are ~100k submitted posts and comments would be used. The dataset would be from subreddits of Windows and MacOS, from January 2021 to the end of October 2022. The text would be combined with the title and body text of the submission and comments.