Team Champagne project proposal

Sentiment Analysis of Google Reviews

ZongYun Li(netID: zyl2), Chi Han(netID: chihan3)

October 23, 2022

Topic Selection Our free topic is Using NLP tools to categorize Google reviews. We found that many users in google reviews may falsely rate 1 or 2 stars, but gave a good comment. So our task is using NLP model to detect whether the rating fit the comment.

Project Plan We are going to use python as our programming language, and use packages such as HuggingFace and PyTorch as analyzing tools. We will use google reviews to be our dataset. The label will be the right or wrong that whether the comment is related to the rating, and the feature will be the comment and the rating. The way we collect our dataset is Crawler, and we decide to use Bert and RoBERTa as our model. We will use cross-validation as our evaluation, and we expect that the accuracy can be higher than 90 percent.

Workload and Division of Labor The workload of this project includes the following:

- Dataset Crawling (∼10 hours) We plan to use data crawling tools such as Crawler to collect reviews and ratings information from the Internet. We will split the dataset for cross-validation.
- Dataset Labelling (∼15 hours) As the labels provided in the dataset may not be accurate (e.g., when the user miss-clicked the button), we will manually label the dataset to create an alternative ground-truth for the dataset.
- Model Tuning (∼15 hours) We plan to build a sentiment analysis tool built upon BERT and RoBERTa. We will use the sentence embedding as representation and train a multi-layer perceptron (MLP) on the dataset.
- Evaluation and Result Analysis (∼10 hours) Finally, we plan to quantitatively evaluate the model's performance on the dataset we collected. We will also conduct analysis of the model's behavior and provide error analysis to provide guidelines for future improvement.