CS 410 Final Project Progress Report

Jiahao Zhang: jiahao4, Mingqing Sun: ms19, Ruining Tao(Captain) : rtao6, Ziyue Zhou: ziyue5

Progress Made So Far: We examined several data sets and we chose one data set[1] published on Kaggle by Arbuzova et. al. The dataset contains over 60k tweets in 2021 and they are categorized into positive, neutral and negative sentiments. We plan to use these labeled data as ground truth. We researched and selected several different python tools which can be used in sentiment analysis, including TextBlob, Vader, TweetTokenizer.

Remaining Tasks: We will use the data collected to examine the performance of each tool mentioned above and select the model with best performance. We will use this model to build a web application that will take the user's sentence as an argument and return the sentiment analysis result of that sentence. We also have to deploy our backend application and build frontend application.

Challenges Encountered: There are many models that can be used in sentiment analysis, and it's challenging to pick the most suitable model for our task: sentiment analysis of tweets. We read several papers in which those models were proposed and it's time-consuming to understand which model should be used under what kind of context. We are rather new to this field and it's also challenging to understand terms such as lexicons etc. in those papers.

1. https://github.com/ConstaT99/410project/tree/main/data