CS 410 Project Proposal

Author/Captain: Rustom Ichhaporia (netID: rustomi2)

Task

I plan to conduct a sentiment analysis on my personal text messages from different platforms. As a baseline, I'd like to be able to identify the sentiment of individual messages. For example, I'd like to be able to identify that the message "I'm so excited to see you!" is positive, while the message "I'm so sad to see you go!" is negative. I also want to be able to identify the sentiment of a conversation as a whole. For example, I'd like to be able to identify that a conversation with Person A tends to be positive, while a conversation with Person B tends to be negative, as well as trends in sentiment over time across all conversations. I will make use of existing pre-trained models as well as my own dataset for the classification task. Evaluation will largely be based on my manual, by-hand classification of a sample of my messages, as there is no existing labeled dataset for this task.

Datasets

I plan to use some combination of my messages from Google Hangouts (accessible via Google Takeout), Apple iMessage, and Facebook Messenger. I may not end up using all three sources based on the difficulty of extracting the individual text information from each source. As a fallback plan, I can connect to the Twitter API and use tweets about text messages as a dataset, which is a more well-documented task.

Tools

I will use online scripts along with SQL to extract the text message data from their respective databases. I will use Python for the data handling and prediction portion of the project, likely in conjunction with the NLTK and HuggingFace libraries. I intend to fine-tune a pubic model using some external labeled data, and then classify the individual texts from my database and aggregate the outputs. The work is exploratory rather than having a definite deliverable goal, so I may alter my approach as I go to produce interesting results.

Time Commitment

I estimate that the workload of this project will be at minimum 20 hours. I will spend 5 hours extracting the data from the different sources, 5 hours cleaning the data, and 10 hours training and testing the model. I will additional time to write the report and prepare the presentation. It is possible that certain steps (data extraction) may take proportionally longer than expected, and I may adjust the complexity of the output accordingly.