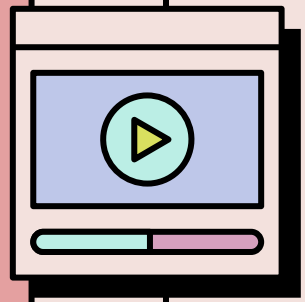
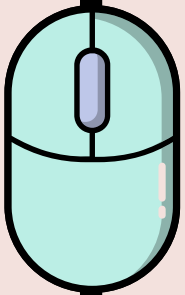
TMI Reddit Content Moderation Week 1





Icebreaker

Introduce yourself! Include:

1. Your name
2. Your major and/or minor(s)
3. What you want to learn/achieve in this team
4. A hobby not related to academics that you love!

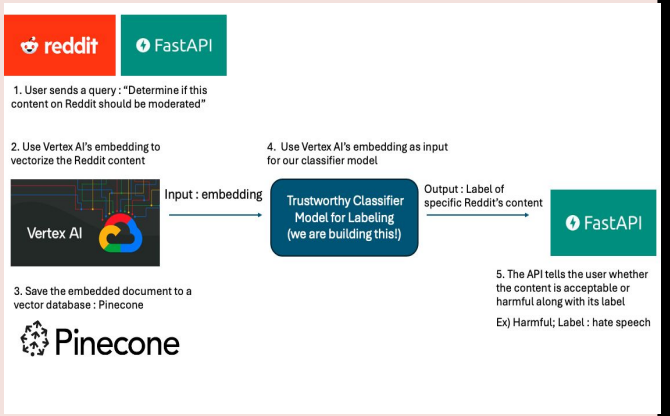
Timeline

October : Understand the project
-Literature Review & Project Scope Definition & Start Data Collection

November : Learn how to do embedding & Store embedding
-Introduction to Google Cloud Platform - Vertex AI embedding tool
-Test runs with small reddit posts
-Plan for how to store embeddings -> Pinecone (Vector Database)

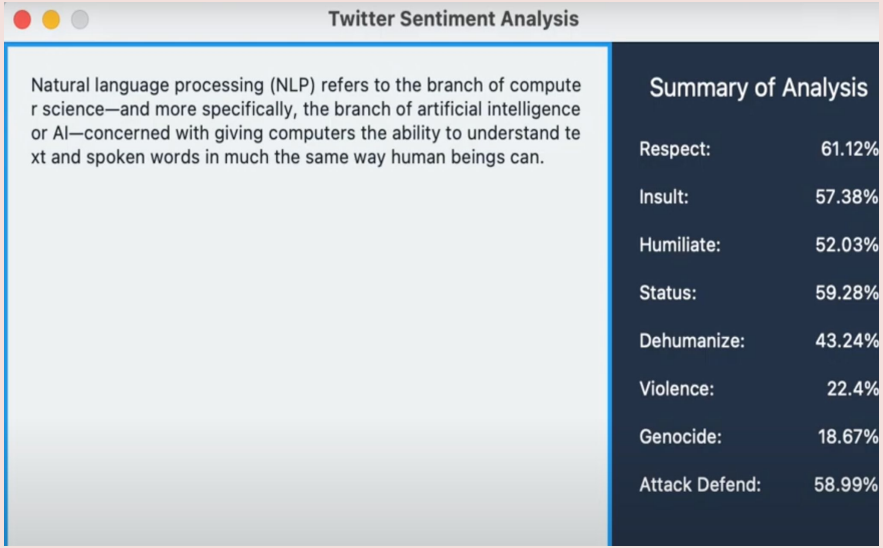
December ~ February : Build classifier model
-Plan for model architecture -> Deep learning based model
-Find statistical tools for classifier model to capture high dimensional embeddings
-Collect labeled data to train classifier model
-Evaluation & Fine tuning

March - April : UI development & Report Writing
-Further fine tuning and advanced prompt engineering if there's time



Previous Project Review

https://epai-sat.w3spaces.com/



Problem: Toxic posts, hate speech, and harmful language.

Dataset: **Measuring Hate Speech Dataset** (39,565 comments); Label : Respect, Insult, Humiliate, Dehumanize, Violence, and Genocide.

Data Processing: Comments were cleaned and tokenized. Created custom word embeddings due to language complexity, split into training (70%), validation (15%), and testing (15%) sets.

Models: We tested RNN, LSTM, and GRU models. GRU achieved the best performance with an F1 score of 60%.

RNN(Recurrent Neural Network) : Recognize pattern in short term dependencies

LSTM(Long Short-Term Memory) : Input, Forget, Output gates

GRU(Gated Recurrent Unit) : Reset and Update Gate