Goals Recap

The main goal of this project is to develop an AI system that can accurately recognize the emotions expressed in KakaoTalk status messages and recommend songs that fit those emotions. The system will suggest personalized music recommendations based on the user's emotional state, as expressed through their profile's status message. To achieve an important part of this goal, we have ascertained that the collection of status messages will be manually collected through means of surveys and questionnaires. Additionally, after deliberation of the previous week's feedback, we have decided to adjust our novelty factor to be a percentage-based emotion classification/matching based on a user's profile status message.

Data Statistics: Datasets

We have selected three different sources to use as our dataset.
From AI-HUB we have sourced data from three datasets, Emotions in Continuous Korean Conversations Dataset, Emotions in Discrete Korean Conversations and Audio of Korean Conversations for Sentiment Analysis Dataset. From the National Institute of Korean Language we have used their 모두의 말뭉치 dataset, and from ETRI Nanum we have sourced from their Emotions in Korean Language Audio Dataset.
However, due to difficulties in acquiring datasets from the National Institute of Korean Language and ETRI Nanum, as of now, we have only sourced data and used the datasets provided by AI-Hub.

Data Statistics: Labels

From the three datasets we have implemented thus far, they all feature seven labels with an insignificant difference between the three. The two text-based datasets both have their labels in Korean being 행복, 슬픔, 놀람, 분노, 공포, 혐오, and 중립. The audio-based dataset has the same respective labels but in English.

Data Statistics: Before Preprocessing

The following are the raw statistics from the compiled dataset that we had before preprocessing.
Entries: 138,212 sentences
Labels: 7
행복: 11,615
슬픔: 21,239
놀람: 12,519
분노: 20,934
공포: 9,697
혐오: 10,309
중립: 51,899
As you can see, there are some issues that just this brief outline of the statistics alone present. We will go over the steps we took to preprocess our data in the next slide.

Data Preprocessing and Normalization

The first step we took in preprocessing was to compile the three datasets we had into one file for simplicity and ease of use. Next, we figured that the labels should all be in one language and thus standardized the language to be Korean. Third, due to the datasets being manually compiled, we encountered some simple misspelling and typos in the labels and corrected them accordingly. Lastly, as shown in the previous slide, the number of sentences for each label ranged from 9,697 being 공포 to 51,899 being 중립. We decided to address this issue by normalizing this amount through a random selection process and ended up with a one-to-one ratio of all the labels.

Data Instances:

In this slide, we have outlined five interesting data instances from our dataset. We picked these five instances to denote the possibility of varied interpretations. Starting off, the first sentence is labeled as fear. When viewed in a simplistic manner, the fact that elevator and fear can be associated is interesting since it's simply a tool or object and should not inherently induce fear. The second sentence also proposes a similar aspect to the first. Mold in of itself should not produce a reaction of anger. Perhaps a disgusted response might be more appropriate, however with the way this sentence is stated it was classified as anger

The third sentence deviates from the word and emotion pairing abnormality and towards a more sentence to emotion pairing abnormality. One could definitely say this with anger, but it seems like this sort of sentence would be stated in a more defeated tone or as a simple observatory tone, which in our case would be classified as neutral. The fourth sentence is classified as a neutral emotion, but it seems apparent that the same sentence could be paired with a happy or an excited emotion. Again, the fifth sentence expresses a similar ambiguity between the sentence and the classified emotion. One could recognize that the usage of the periods could present a tone of anger which is commonly understood as such by the younger generation. However, an older generation might interpret this simply as a reply or answer to a question or statement thus classifying this sentence as neutral.

Baseline Model

We picked KoELECTRA, the Korean Language variant of ELECTRA for our baseline model. ELECTRA utilizes a generator model to replace masked tokens in a sentence while a discriminator model is trained to distinguish between the original and replaced tokens. This approach of training performs more efficiently compared to MLM. We chose to use KoELECTRA as our baseline model because of its state-of-the-art performance in Korean Sentiment Analysis which aligns perfectly to our task.