Project Proposal : Predicting Emotions from Speech Signals And Text

2277044, leeseunghyeon
2491007, kimyeeun
2491021, leebada
2491036, jutaein

1.Business Value

1.1 Background

The rapid spread of generative AI technologies such as ChatGPT is fundamentally transforming the way humans interact with machines. Users are increasingly engaging with AI systems through everyday queries and emotional expressions. However, these systems still face limitations in accurately recognizing users' emotional states. Given this limitation, there is a growing demand for more sophisticated emotion recognition technologies capable of capturing affective cues beyond textual content.

Moreover, many are primarily designed with English-speaking users in mind, resulting in additional challenges for Korean users due to language and cultural gaps. Our approach seeks to explore the potential of detecting emotional states by actively leveraging nonverbal speech signals alongside textual cues.

1.2 Problem Definition & Value Proposition

Existing research on emotion recognition has largely relied on long-form text interviews, which often fail to capture subtle emotional shifts in everyday contexts. This project aims to develop a machine learning model designed for Korean users, capable of detecting emotional states using short, natural speech recordings.

Speech conveys emotion effectively through nonverbal features such as tempo, emphasis, and vocal tremors, making it particularly useful in the field of emotion recognition. By applying a model that integrates these speech signals with a text-based classifier, it is expected to maximize emotion classification performance. It not only helps overcome the limitations of conventional approaches, but also offers a wide range of applications, including integration with wearable devices, real-time emotion prediction, and personalized feedback delivery.

2. Available Data

2.1 Data Acquisition

This project utilizes the 5th-year, 2nd-phase emotion classification dataset ("감정분류를 위한 데이터셋") available on AIHub. The dataset was developed by the KAIST Artificial Intelligence Research Institute. It consists of 19,374 Korean audio files collected through an emotional dialogue application. Each file is labeled with one of seven emotions(happiness, anger, disgust, fear, neutral, sadness, and surprise)and was annotated by five raters. A provided CSV file contains additional metadata, including the selected five emotions, their intensity levels, and the transcribed utterances.

2.2 Data Review

The data imbalance across emotion categories was addressed using a downsampling approach. The restructured dataset contains 17,531 samples, including the primary emotion label, the utterance text, and acoustic features extracted through openSMILE.

To assign a single primary label, the five individual annotations were integrated by aggregating scores for identical emotions, applying weights based on intensity. Only the emotion with the highest weighted score was retained as the final label. Additionally, the 'neutral' label was excluded from the dataset.

3. Formulation

3.1 Choice of Algorithms and Rationale

This project uses a total of 17,000 samples and 62 audio features. In this setting, fast convergence and prevention of overfitting are critical. The CatBoost Classifier meets these requirements by automatically handling categorical variables, incorporating built-in regularization techniques, and providing GPU support. Therefore, we have adopted the CatBoost Classifier as the primary classification model for this project.

3.2 Input & Output Variables

The input variables for the prediction model include the utterances and 62 audio features(loudness, amplitude, frequency, etc.) extracted using eGeMAPS from OpenSMILE library. The output variable is the predicted category of emotions.

3.3 Expected Challenges and Alternatives

In this study, text features were extracted using the Okt morphological analyzer and TF-IDF. However, due to the limitations of handling compound nouns in Korean, there is a risk of semantic distortion or information loss.

As a potential solution, more accurate analyzers such as Mecab or context-aware embedding methods based on BERT can be considered. Additionally, the early fusion approach—which simply concatenates audio and text features—may lead to performance degradation due to differences in feature scales. This issue can be mitigated by applying feature normalization or adopting a late fusion strategy.

Class imbalance in emotion labels may also cause the model to be biased toward certain classes. To address this, techniques such as class weighting or data augmentation will be applied.

Finally, to improve model generalization, feature selection methods will be employed, and their impact on performance will be systematically evaluated.