CS626: Project Discussion
Emotion detection from audio and text

Deepak Singh Baghel 203050005
Ankush Agarwal 203050007
Nilesh Kshirsagar 203059004
…

27th November, 2021

Motivation

- Emotion analysis is useful in identifying how users feel about a product based on reviews
- It is used in healthcare domain to identify and monitor certain conditions such as stress, anxiety and depression
- It is also useful in development of virtual assistants

Problem Statement

- Problem Statement

  To identify the emotion of speakers in a conversation based on audio and text modalities

  – Input : Audio file and corresponding transcript

  – Output : Emotion(Sad, happy, neutral etc.)

-

Supporting Papers (Literature survey)

- Basic

  Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text. In 2018 IEEE Spoken Language Technology Workshop (SLT) 2018 Dec 18 (pp. 112-118). IEEE.

- State-of-the-art

  Siriwardhana S, Reis A, Weerasekera R, Nanayakkara S. Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. arXiv preprint arXiv:2008.06682. 2020 Aug 15.

- Implemented

  Sahu G. Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022. 2019 Apr 12.

Data

IEMOCAP Dataset

It contains five recorded sessions of conversations from ten speakers and amounts to nearly 12 hours of audio-text information along with transcriptions.

It is annotated with eight categorical emotion labels, namely, anger, happiness, sadness, neutral, surprise, fear, frustration and excited.

Main Techniques

- Data Preprocessing
  - Audio features - pitch(autocorrelation), harmonic(median), speech energy(rmse), Pause(threshold compared to rmse), Central Moments(mean and standard deviation).
  - Text features - BERT encoding using bert-base-uncased model
- Concatenated audio and text features.
- Combined labels: happy and excited, sad and frustrated.
- Classical ML techniques: Random Forest (RF) , Gradient Boosting (XGB) , Multinomial Naive Bayes (MNB) , Logistic Regression (LR)

Continued...

- DL techniques: Multi Layer Perceptron (MLP) , Bert based uncased model.
- Used Bert encoding instead of TF-IDF vectorizer used in the paper which improved accuracy
- Ensembled (RF + XGB + MNB + LR + MLP)
- For analysis used audio files as test set, converted audio to text data using asr library. Extracted features from text and audio using above techniques and predicted the emotions.

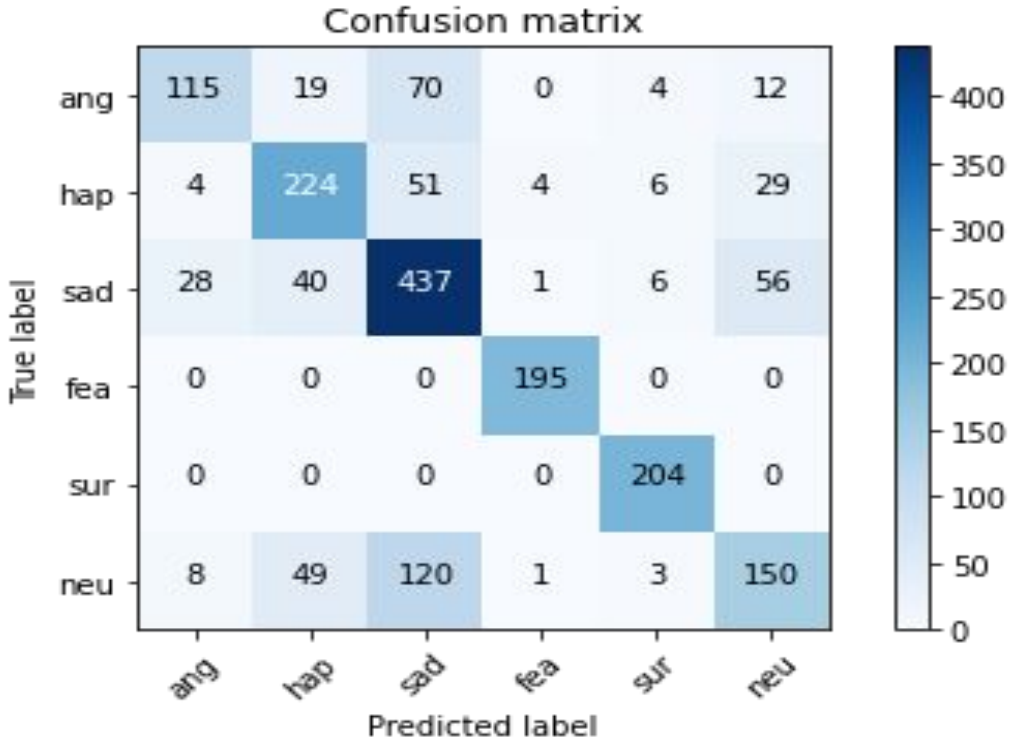
Results

Metrics for Ensembled Model (RF + XGB + MNB + LR + MLP)	Values (in %)
Accuracy	72.2
F1 score	74.4
Recall	74.2
Precision	75.9

Analysis: Sentence Emotion Predicted

1. That was a good one, for a second I was like ohhh. Happy
2. Today the weather is not good. What happened? Neutral

Confusion Matrix



Conclusion

- In this project, we tackled the task of speech emotion recognition and study the contribution of different modalities on the IEMOCAP dataset.
- Used both ML and DL based models.
- We have seen that ensembling multiple ML models leads to improvement in the performance.

Demo