**Ahmad Tabsho** **200504072**

# Multimodal Sentiment Analysis

## Abstract:

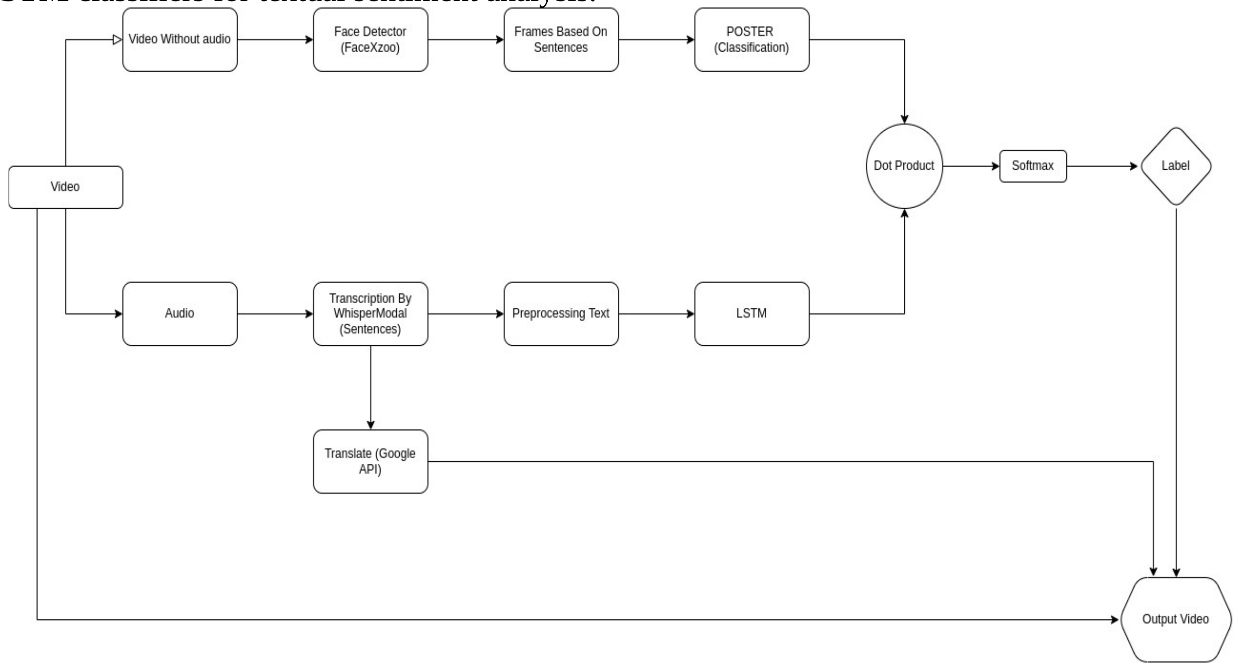
This report explores a multimodal sentiment analysis project that integrates data from both video and audio to assess sentiment more accurately than unimodal methods. The project involves detecting faces in video frames and classifying sentiment using the POSTER model, while concurrently transcribing audio using WhisperModel, preprocessing the text, and classifying sentiment using an LSTM classifier. The results from both classifiers are combined through a softmax-based dot product to determine the overall sentiment. The final sentiment and text, translated into Turkish, are displayed with subtitles. The findings demonstrate the effectiveness of this approach in providing nuanced sentiment analysis for multimedia content. **Code Sources [https://drive.google.com/drive/folders/1T1XMsrnmHWzYkQO-ipRj5Y-_fuO5dvc-?usp=drive_link]**

## Introduction:

Sentiment analysis, or opinion mining, involves determining the sentiment expressed in text, speech, or visual content. Traditional methods often rely solely on text, which limits their accuracy due to the lack of contextual and non-verbal cues. This is particularly challenging in understanding nuanced human emotions that are often conveyed through facial expressions, tone of voice, and body language.

Multimodal sentiment analysis aims to overcome these limitations by incorporating data from multiple sources. By analyzing text, audio, and visual inputs simultaneously, multimodal approaches provide a richer and more accurate understanding of sentiment. This is especially important in applications such as social media monitoring, customer feedback analysis, and human-computer interaction, where understanding the full spectrum of human emotions can lead to better insights and more effective responses.

The technogloies that I have used:
**FaceXzoo** for face detection in video frames,
**POSTER** for visual sentiment classification,
**WhisperModel** for audio transcription,
**LSTM** classifiers for textual sentiment analysis.



**Modal Diagram**

**Methodology:**

The project involves several key steps, combining video and audio processing to achieve accurate sentiment analysis:

**1. Data Collection:** Obtain a video containing spoken sentences. This video serves as the input for both video and audio processing pipelines.

**2. Video Processing:**
**- Face Detection:** Use FaceXzoo to detect faces in the video frames. Face detection helps isolate the parts of the video where facial expressions are present, which are crucial for visual sentiment analysis.
**- Frame Extraction:** Save frames corresponding to each spoken sentence. For instance, if there are four sentences in the video, save four frames. This ensures that each frame is aligned with a specific spoken segment.
**- Sentiment Classification:** Feed the saved frames into the POSTER sentiment classifier. POSTER analyzes facial expressions and other visual cues to predict the sentiment for each frame. POSTER is trained on the RAFDB dataset, which consists of images labeled with basic emotions, including 'Surprise', 'Fear', 'Disgust', 'Happiness', 'Sadness', 'Anger', and 'Neutral'. This dataset provides a diverse range of facial expressions for robust sentiment analysis.

**3. Audio Processing:**
**- Transcription:** Use WhisperModel to transcribe the audio portion of the video into text. Accurate transcription is critical for subsequent text-based sentiment analysis.
**- Text Preprocessing:** Apply preprocessing operations, such as removing stop words, normalizing text, and handling punctuation, to clean and prepare the text for analysis.
- Sentiment Classification: Provide the preprocessed text embeddings to an LSTM classifier. The LSTM model analyzes the sequence of words to predict the sentiment of the spoken content. The LSTM classifier is trained on a dataset of text pieces, such as tweets, labeled with emotions including 'Surprise', 'Fear', 'Disgust', 'Happiness', 'Sadness', 'Anger', and 'Neutral'. This dataset is selected to align with the sentiment labels used in the dataset that collected from tweets.

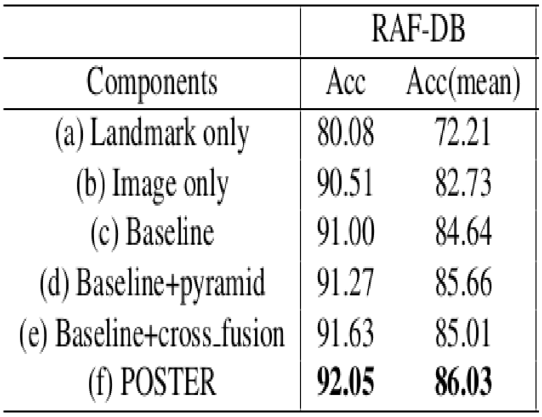
**4. Combining Modal Results:**
**- Softmax Application:** Apply softmax to the sentiment predictions from both the video (POSTER) and audio (LSTM) classifiers to obtain probability distributions.
**- Dot Product:** Compute the dot product of the two softmax results to integrate the predictions from both modalities.
**- Final Sentiment Decision:** Apply softmax again to the combined result to determine the final sentiment.

**5. Translation and Display:**
**- Translation:** Translate the English text of the audio transcription into Turkish using the Google Translate API accessed through the `googletrans` library. This API provides reliable translation services for various languages, including Turkish.

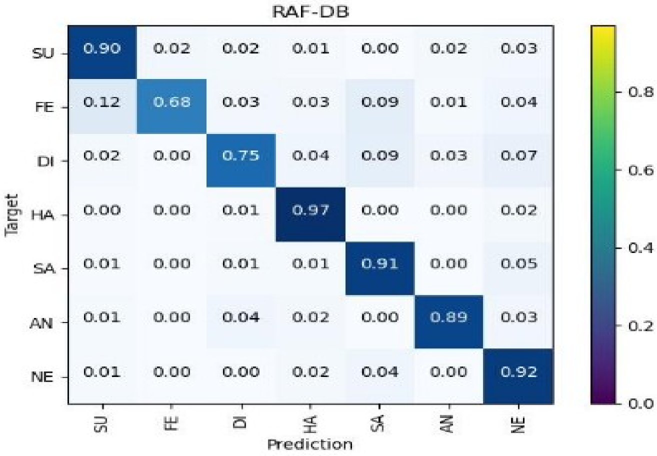
**- Subtitle Display:** Show the sentiment for each sentence or section of the video along with the translated subtitles.

**Results:**

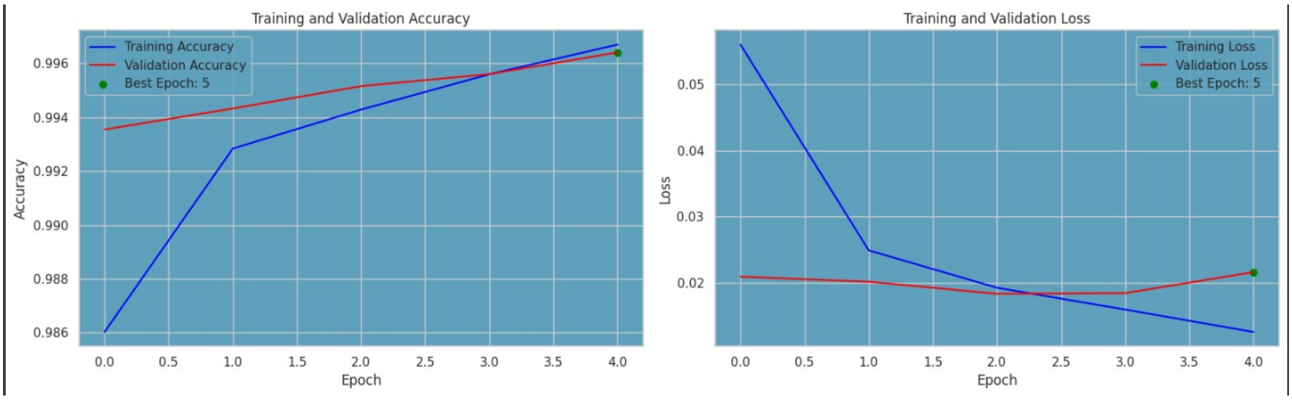
The results of the visual modal(POSTER) :

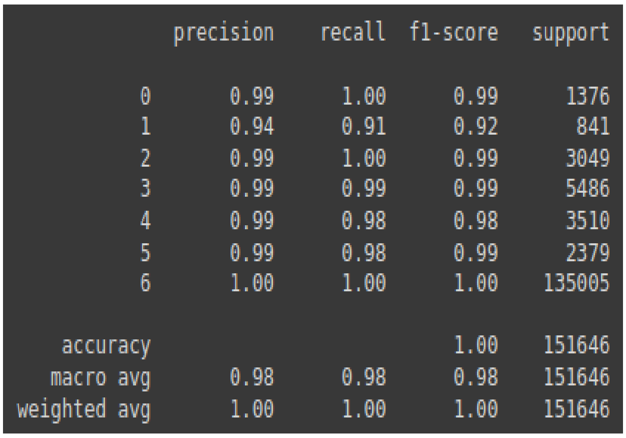


| | RAF-DB | |
|---|---|---|
| Components | Acc | Acc(mean) |
| (a) Landmark only | 80.08 | 72.21 |
| (b) Image only | 90.51 | 82.73 |
| (c) Baseline | 91.00 | 84.64 |
| (d) Baseline+pyramid | 91.27 | 85.66 |
| (e) Baseline+cross_fusion | 91.63 | 85.01 |
| (f) POSTER | **92.05** | **86.03** |



The results of the text classifier modal(LSTM Classifier):

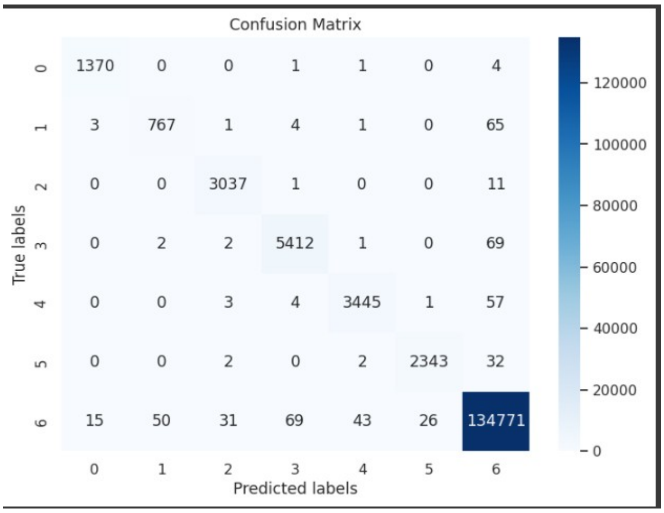




```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1376
           1       0.94      0.91      0.92       841
           2       0.99      1.00      0.99      3049
           3       0.99      0.99      0.99      5486
           4       0.99      0.98      0.98      3510
           5       0.99      0.98      0.99      2379
           6       1.00      1.00      1.00    135005

    accuracy                           1.00    151646
   macro avg       0.98      0.98      0.98    151646
weighted avg       1.00      1.00      1.00    151646
```



## Conclusion:

Multimodal sentiment analysis provides a significant improvement over traditional unimodal methods by leveraging the strengths of various data modalities. This approach allows for a more accurate and nuanced understanding of sentiment, which is vital for applications in social media monitoring, customer feedback analysis, and human-computer interaction.

By integrating information from video, and text, multimodal analysis captures a more comprehensive picture of sentiment, accounting for contextual nuances. This deeper understanding enables better insights into public opinion, brand perception, and user sentiment across various platforms.

In conclusion, multimodal sentiment analysis holds promise for enhancing decision-making processes and improving user experiences across a wide range of domains. As research in this area continues to evolve, we can expect further advancements that will strengthen its effectiveness and applicability in real-world scenarios.

## References:

1. Cai, Y., Zhang, H., Zhang, Y., & Tian, Y. (2023). MARLIN: Masked Autoencoder for Facial Video Representation LearnINg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [https://openaccess.thecvf.com/content/CVPR2023/papers/Cai_MARLIN_Masked_Autoencoder_for_Facial_Video_Representation_LearnINg_CVPR_2023_paper.pdf]

2. Zheng, C., Mendieta, M., & Chen, C. (Year). POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition. Center for Research in Computer Vision, University of Central Florida. [https://arxiv.org/pdf/2204.04083]

3. Wang, J., Liu, Y., Hu, Y., Shi, H., & Mei, T. (Year). FaceX-Zoo: A PyTorch Toolbox for Face Recognition. JD AI Research, Beijing, China. [https://arxiv.org/pdf/2101.04407]