Received:24/04/2024 Revised: 26/08/2024 Accepted:25/09/2024

Journal of Emerging Computer Research and Applications (ISSN: 3048-7773)
Volume 1/Issue 2/July-Dec 2024
Research Article

Multimodal Emotion Recognition

Jeet Manjrekar[1] , Vijesh Mundokalam[2] , Devesh Sali[3] , Sakshi Rajeshirke[4], Prathamesh Mane[5]
[1, 2, 3, 4,5]A. P. Shah Institute of Technology, Kasarvadavali, Thane, India
Email Id:letsmailjeet@gmail.com, vmundokalam@apsit.edu.in, dvesh.sali28@gmail.com ,
sakshirajeshirke89@gmail.com, maneprathamesh78@gmail.com
Corresponding Author ORCID: 0009-0007-9470-5835

Abstract

This study investigated multimodal emotion recognition frameworks that combine information from various sources, such as facial expressions, speech patterns, and physiological signals, to comprehensively understand human emotions. It reviews advancements in machine learning and deep learning models, and their applications in real-world scenarios. The challenges and opportunities related to the implementation of multimodal emotion recognition frameworks were also discussed. The field of emotion recognition has applications in various domains, including healthcare, marketing, and human-computer interaction. Technology can improve human-machine interaction and revolutionize the way we interact with technology. The abstract underscores the importance of continued research aimed at developing accurate and ethically sound multimodal emotion recognition systems. [i]

Keywords: Human-Computer Interaction, Multimodal Emotion Recognition, Affective Computing, Emotion Recognition, Natural language processing.

1. Introduction

Translating feelings in human composing requires a high level of expertise. Content mining investigation right now exists in the areas of characteristic dialect handling Natural Language Processing (NLP) and machine learning (ML). Since the foundation thought of the story and certain words in the content play a vital part in understanding the content. Inquire about enthusiastic investigation has extended to incorporate an assortment of applications; for case, from enthusiastic investigation of information investigation to the improvement of passionate communication chatbots. [1]

Facial acknowledgment, counting facial expressions and feelings, has been a point of intrigue to computer researchers for more than a decade. Recognizing and understanding individual feelings can be vital in numerous ranges, such as open security, healthcare (counting adjustments), or excitement. Besides, the capacity of today's machines to recognize and express feelings will overcome impediments to "characteristic" intuition between machines and people. Common highlight extraction procedures consolidate Mel-frequency Cepstral Coefficients (MFCCs), terrible highlights, and truthful descriptors. [2]

Once highlights are removed, classification calculations are utilized to recognize sentiments from sound signals. Coordinated machine learning calculations such as Support Vector Machines (SVM), k-nearest neighbors (k-NN), and Sporadic Timberlands are commonly utilized for classification assignments. Too, significant learning plans such as Convolutional Neural Networks (CNNs) and Monotonous Neural Frameworks like Recurrent Neural Networks (RNNs) have shown promising results in capturing complex plans and associations in sound information. So distant, programmed discourse acknowledgment (ASER) is a valuable inquiry in the field of Human-computer interaction (HCI) and has a wide run of applications. For case, in e-learning, the computer can analyze the subject or the person's disposition and alter the substance of the subjects or male-female understudies. It is utilized for the convenient examination of complaints in programmed inaccessible calls. In

[1] Corresponding Author: Jeet Manjrekar, A. P. Shah Institute of Technology, Kasarvadavali, Thane, India. Email Id: letsmailjeet@gmail.com, Orcid: 0009-0007-9470-5835

[2] Cite As: Jeet Manjrekar,Vijesh Mundokalam, Devesh Sali , Sakshi Rajeshirke and Prathamesh Mane(2024), Multimodal Emotion Recognition, *Journal of Emerging Computer Research and Applications*, 1(2), 01-12.

mechanical technology, human uneasiness can be recognized by instructing robots to respond to people and choose upon human feelings. Indeed in a therapeutic setting, look at the patient's readiness to experience a mental well-being examination [3]

The classic programmed discourse machine is less known due to a few of the straightforward non-linguistics passed on through discourse, such as sex, identity, feeling, brand target, and heart. The human intellect employments all talked dialect and extra-linguistic data to get discourse – its inactive importance and great spelling. It centers especially on recognizing energetic states from the talked tongue. Sentiments appear in talk through diverse acoustic highlights, prosodic prompts, and phonetic substance. Common acoustic highlights utilized for talk feeling acknowledgment consolidate pitch, imperativeness, formant frequencies, and prosodic highlights such as term, pitch frame, and raised varieties. [4]

Similar to sound feeling acknowledgment, extraction is taken after by classification utilizing machine-learning or deep-learning calculations. In any case, talk feeling acknowledgment stances one-of-a-kind challenges such as speaker capriciousness, phonetic substance, and pertinent components. Investigators routinely utilize methods such as speaker normalization, counting assurance, and setting modeling to address these challenges and move forward with acknowledgment precision.

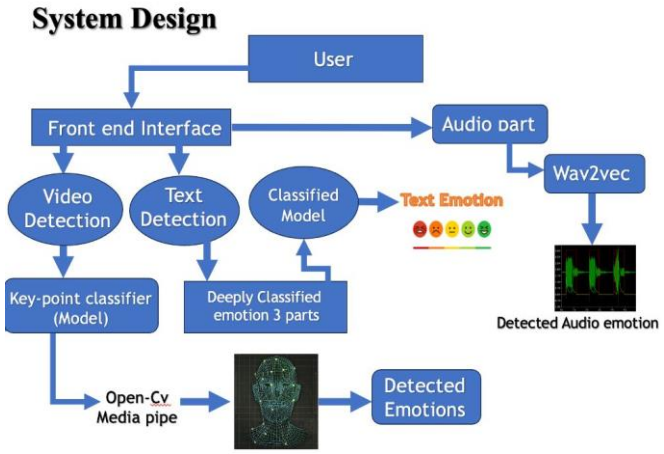

Figure 1. Proposed System Design

2. Literature Review

Programmed human emotion recognition is one of the most critical and developing areas of inquiry in the human-computer interaction (HCI) space. It has a noteworthy effect on applications such as programmed human behavior investigation and mixed media recovery frameworks. As of late, Profound Neural Systems have accomplished impressive victories in terms of exactness related to machine-learning assignments. Programmed human feeling acknowledgment has also been accomplished utilizing a profound learning-based convolutional neural arrangement (CNN). Right now, most profound learning-based calculations for human emotion recognition center as it were on specific headings, such as Vision, Content, and Sound. These calculations were prepared on a particular methodology (visual, printed, and acoustic information) and performed well in a controlled environment. Be that as it may, they fizzled to accomplish great things in most real-life cases. This is because of the erratic behavior of human nature. So to handle this issue, a novel profound learning-based multi-modal design has been proposed in this paper. [5]

This is one of the ways to accomplish this is by comparing chosen facial highlights from the picture and a facial database. Recognizing feelings from pictures has ended up a dynamic investigative topic in picture handling and applications based on human-computer interaction. This ponder conducted a test examination on recognizing facial feelings. The stream of our emotion recognition framework incorporates the fundamental forms of the FER framework. These incorporate picture procurement, preprocessing of a picture, confront location, extraction, and classification, and at that point, when feelings are classified, the framework allocates the client-specific music concurring with their feelings. The all-around acknowledged feelings considered for the tests included joy, Pity, Astonish, Fear, Appall, and Outrage. [6] To begin with, hand motions for sign dialect acknowledgment and facial feelings are prepared utilizing a Convolutional Neural arrangement (CNN), and at

that point by preparing the emotion-to-speech demonstration. At last, hand motions and facial feelings were combined to realize feelings and discourse. [7]

Automatic recognition of emotions from human speech is currently a widely researched topic. This article attempts to analyze and classify speech thought in three languages: Berlin, Japan, and Thai. Speech features such as frequency (F0), power, zero drift rate (ZCR), line projection (LPC) and Mel frequency cepstrum coefficient (MFCC) of short-duration wavelet signals were obtained by graduation. In this case, Support Vector Machine (SVM) is used as the classification model. [8] Confront discovery is an critical issue in computer vision. Key examination is done by evaluating the arranges of a few faces. In this paper, convolutional neural systems are utilized to appraise facial keypoint discovery. This demonstrate is prepared to appraise facial region utilizing webcam input. The fundamental highlights of the confront incorporate the corners of the eyebrows, the tip of the nose, the corners and regions of the eyes, and lip subtle elements. The anticipated keypoints are coordinated with the webcam input and compared with diverse models for a way better understanding of the substance. The cruel square mistake is utilized to assess the misfortune of each test. [9]

Speech recognition (SER) is an important area of research in collaborative and social robots to improve human-robot relationships (HRI) and is a suggestion for thinking about thinking. Despite recent advances in SER research, the problem is still difficult to investigate due to the complexity, content, and large differences in different states of the human heart. Therefore, the problem that arises in the design of the language of paralinguistic thinking in description is more serious when working on educational monitoring, because it requires writing for documents large enough to achieve high performance standards. To this end, self-directed learning (SSL) is widely used in the language field to solve the problem of limited data availability. [10]

3. Model

This segment depicts the strategies utilized for the discourse recognition assignment. To begin with, we show a reproducible encoder show for audio and content designs. We at that point propose a distinctive strategy that employments a two-loop encoder to encode sound and content at the same time.

3.1. Keypoint Classifier (KPC)

The keypoint classifier for video emotion recognition is a key component in frameworks planned to identify and classify feelings in genuine video streams. Here are the points of interest of the fundamental points:

1. Key confront discovery: This step includes distinguishing and situating key faces in each outline of the live video. These characters frequently incorporate components such as eye corners, and nose and mouth corners. Numerous strategies exist for confront discovery, counting conventional strategies such as inactive modeling ASM (Active Shape Models) or more advanced strategies based on profound learning such as convolutional neural systems (CNN) or confront location calculations (such as DLIB or MTCNN). MTCNN (Multi-Task Convolutional Neural Network) is used to detect facial boundaries with minimal artifact. [15]

2. Highlight extraction: When the primary focuses of the confront are found, the fundamental highlights that will speak to the confront well are extricated from these focuses. Highlight extraction strategies may incorporate calculating the separation between key focuses, measuring the point shaped by key focuses, giving tasteful data around key focuses, or indeed utilizing profound learning strategies to recognize anomalies.

3. Classification: The extricated highlights are at that point set into classifiers to anticipate the feeling related to the confrontation. Numerous machine learning calculations can be utilized for classification, counting profound learning models such as back vector machine (SVM), irregular timberlands, k-nearest neighbors (k-NN), and convolutional neural systems (CNN). The lesson was prepared on a list of information in which facial expressions were related to particular images (e.g., upbeat, pitiful, and irate). While preparing, the show learns to relegate input highlights to passionate names. [8]

4. Real-time handling: Real-time preparing is vital for real-time video applications where moo inactivity and tall throughput are vital. High-performance calculations and optimizations are utilized to guarantee that the fundamental substance, extraction, and classification steps can be completed inside the time constraint of the real-time video stream. Procedures such as parallelization, optimization of the input calculation, and quantization models can be utilized to progress the execution of the scheduler.

5. Integration into enthusiastic insights: The fundamental point of arrangement is integration into broader passionate insights, which regularly incorporates things like video capture, pre-processing, post-processing, and client interface. The classifier takes video outlines from the input stream, forms them by recognizing values and classification levels, and yields the anticipated content. Depending on the particular needs of the application, the yield can be handled or performed as needed.

Given a set of input features X, the model predicts the probability of each class label Y.

$$P(A \ / \ B) = \frac{1}{1 + e^{-(w^T X + b)}}$$

The above expression represents a logistic regression model, where:

Mathematically, P (A|B) is the probability of class A given input B. e is the base of the natural logarithm. W is a vector of weights. X is a vector of input features. b is the bias term.

The formula calculates the probability of the output being Y given the input features X, by taking the dot product of the weights W and the input features X, adding the bias term b, and applying the sigmoid function $1/1+e-z$, where z is the result of the dot product and bias term. This function squashes the output to the range [0, 1], representing a probability

In common, key focuses in key focuses play a critical part in video emotion recognition by recognizing facial highlights, expelling highlights, and fragmenting them precisely. Considerations on real-time video spilling. Its execution specifically influences the viability and appropriateness of enthusiastic insights in an assortment of applications, including human-computer interaction, healthcare, and innovation. [9]

3.2. Transformers

Transformers are revolutionizing many areas of artificial intelligence, including natural language processing (NLP), and can be translated into speech recognition. Here we explain in detail how to use Transformer for the SER experience. [10]

1. Transformers Overview: Transformers is a deep learning model introduced for natural language processing. They have a unique architecture based on a self-improvement mechanism that allows them to capture long-term expectations in the data set.

Unlike traditional Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), Transformer does not rely on sequential data. Instead, they process data input in parallel, making them suitable for capturing complex patterns in coherent and sequential data.

2. Adaptation for speech recognition: To adapt for speech recognition, techniques such as Mel Frequency Cepstral Coefficients (MFCC) or spectrogram are often used to convert the speech input signal into a vector symbol. Such systems are fed into the transformation model, which allows them to use self-monitoring techniques to capture the progression of different parts of the speech signal. Transformer models can be enhanced with additional layers or adjustments to better capture emotions in speech data. For example, additional layers of position coding can be added to provide the model with information about the physical order of ideas.

3. Training and improvement: The Transformer model is trained on a recording of speech patterns, where each pattern is associated with an emotional expression (such as happy, sad, or angry). During training, the model learns to describe the content of the conversation into the emotional map by reducing loss (such as categorical cross-entropy). Additionally, trained Transformer models such as BERT (Bidirectional Encoder Represented by Transformers) or Generative Pre-trained Transformer (GPT) can be optimized on small datasets, write these recommendations specifically for this purpose. [11, 12]

4. Inference and prediction: Once the Transformer model is trained, it can be used to make inferences about new unseen speech patterns. During inference, the model uses conceptual speech features and a map learned from training to predict emotional symbols. Emotion tags can be further processed or used directly in documents, such as in emotion-sensitive human-computer interaction or emotional intelligence in customer service.

5. Model evaluation and performance: Evaluate the performance of the Speech Satisfaction Transformer model on separate datasets using metrics such as accuracy, precision, recall, and F1 score. Additionally, qualitative analyses such as error analysis and weight monitoring can provide insight into how the model makes predictions and the extent of input is related to emotional awareness.

Application of Transformers:
Transformer is a neural network with great potential in many areas, including healthcare. When it comes to audio use in the medical field, transformers can be used in a variety of ways:

1. Speech Recognition: Transformers can be used to accurately record medical conversations, patient interviews, or doctor-patient conversations. This can improve the clinical process, reduce errors, and increase doctors' productivity.

2. Automatic Medical Transcription: In a medical setting, it is a collection of patient consultations or medical procedures and the Transformer can save records recorded in books. This saves doctors time and increases the accessibility of medical information.

3. Clinical Decision Support System: Transformers can help doctors make medical decisions by analyzing audio data from patient exams or monitoring devices. For example, they can analyze cough sounds to diagnose respiratory diseases or check heart rhythm abnormalities for early detection of heart disease.

4. Monitor patient health: Audio data such as breathing patterns, snoring, or voice characteristics can provide information that the patient's healthy drink is clean. Transformers can instantly process this information to monitor patients remotely, detect abnormalities, and alert doctors in emergencies.

5. Healthcare Chatbots: Transformers can use powerful healthcare chatbots that interact with patients through voice interaction. These chatbots can provide medical advice, answer questions, schedule appointments, and even perform preliminary tests based on symptoms patients describe.

6. Rehabilitation and Treatment: Voice feedback is very important during the rehabilitation and treatment process. Transformers can analyze speech patterns or vocal cues to provide instant feedback to patients receiving speech therapy, cognitive rehabilitation, or mental health counseling.

7. Medical Image Annotation: Although not audio-related, the Transformer can also assist with medical image annotation tasks by linking the content of the annotation or information treatment, thus improving image analysis and diagnostic accuracy.

Overall, Transformer provides a strong foundation for speech recognition, making it a great model for capturing complex and progressive patterns in speech products. With proper training and tuning, the adaptive electronic model can achieve state-of-the-art performance in speech recognition.

3.3. Text to sentiment

The Twitter-emotion-classification-with-bert model predicts the sentiment of a given text (positive, negative, and neutral).

It typically uses a recurrent neural network (RNN), Long Short-Term Memory (LSTM), or Transformer architecture. The model learns to map input text sequences to sentiment labels. Mathematically, the sentiment prediction can be represented as a softmax function:

$$P(A = y_i|B) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z}j}$$

Where B is the input text, A is the sentiment label, yi is the sentiment class (positive, negative, neutral), zi is the logit corresponding to class yi, and N is the total number of sentiment classes.

4. Experimental Setup and Dataset

4.1. Dataset

In our project, we utilized a keypoint classifier dataset as a foundational component for training and validating our machine learning model. This dataset comprises annotated images where key points, such as corners, edges, or specific features, have been identified and labeled. By leveraging this dataset, we aimed to develop a robust classifier capable of accurately detecting and classifying these key points within new, unseen images. This involved preprocessing the data, selecting appropriate features, and employing various machine learning algorithms or deep learning architectures to train the classifier. Additionally, we likely performed rigorous evaluations and fine-tuning to optimize the model's performance, ensuring its effectiveness in real-world applications. The model used in this study is the Twitter Emotion Classification dataset obtained from the Hugging Face repository titled "twitter-emotion-classification-with-bert." It comprises Twitter posts annotated with emotion labels, providing diverse examples for training emotion classification models. The dataset is partitioned into training, validation, and test sets and undergoes preprocessing to enhance model training and evaluation. Here we have done model training on the text emotion dataset to identify various emotions as well as sentiments. Below you can see graphs of emotions and sentiments.

Figure 2. Classification of emotions based on categories of emotion

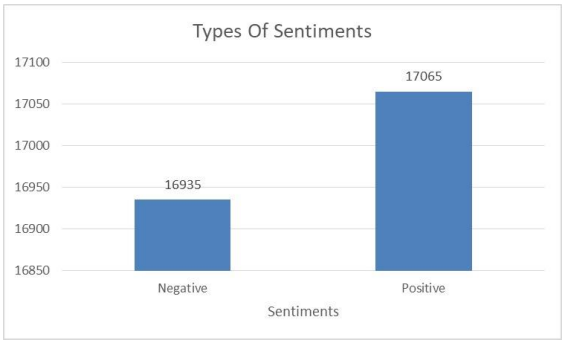

Figure 3. Classification of sentiments based on the type of sentiment

The table below shows the accuracy of the model that we have used for text-to-sentiment analysis.

Table 1: Training Accuracy of Classifier Used

Classifier	Training Accuracy
SVM	72.45%

We tried all the algorithms and finally, we concluded with the SVM algorithm and also compared it with the Kaggle dataset. our dataset is stronger than the Kaggle dataset, because we classified the emotions into three parts i.e. very, slightly, and also added the sentiments i.e. positive and negative and if we delete the sentiment column then we can achieve accuracy up to 90%.

4.2. Feature extraction

For our multimodal emotion recognition research, we employ a combination of feature extraction techniques tailored to the unique characteristics of our dataset and input modalities. The feature extraction pipeline encompasses both facial landmark detection from video frames and audio waveform processing from microphone recordings.

Facial Landmark Detection: We utilize MediaPipe's FaceMesh model for real-time detection of key facial landmarks from video frames captured by the camera. These landmarks, including the eyes, nose, mouth, and jawline, serve as crucial indicators of facial expressions. [13]

Audio Waveform Processing: The audio waveform is captured from the microphone using the PyAudio library, recording audio data in byte format. We convert the raw audio byte string into a tensor representation suitable for processing by machine learning models. This conversion involves normalization and encoding of the audio waveform, facilitating subsequent analysis and transcription tasks.

Integration and Utilization: Extracted facial landmarks and audio features serve as input for our multimodal emotion recognition models. The facial landmark coordinates and audio feature tensors are integrated with

trained models such as Wav2Vec2 for CTC (Connectionist Temporal Classification) and sentiment analysis pipelines. These features enable comprehensive analysis of facial expressions and audio cues, contributing to the accurate classification and understanding of human emotions. [14]

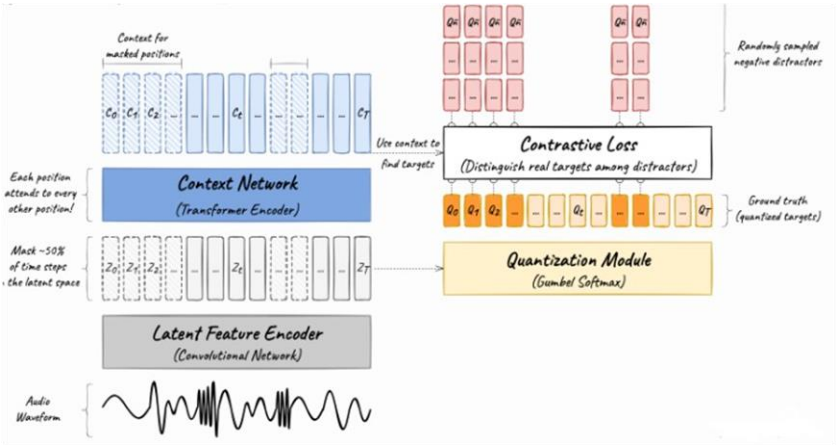

Figure 4. Architecture of Wav2Vec2 [16]

4.3. Implementation details

Our multimodal emotion recognition system integrates facial and audio modalities, enabling real-time analysis of emotional states. The implementation revolves around the selection and integration of appropriate models, preprocessing of input data, real-time processing, user interface development, deployment, and integration with external services.

In the realm of facial emotion recognition, we rely on the MediaPipe library's FaceMesh model for detecting facial landmarks. These landmarks serve as the basis for our custom keypoint classifier model, facilitating emotion inference based on facial expressions. For audio emotion recognition, we leverage the Facebook Wav2Vec2 model trained on audio data for speech-to-text transcription. Subsequently, sentiment analysis is performed using a trained sentiment analysis model from the Transformers library. [14]

The preprocessing stage plays a vital role in ensuring consistency and efficiency in subsequent emotion inference. Facial landmarks extracted by the FaceMesh model undergo preprocessing to obtain a compact and normalized feature representation. Similarly, audio data captured from the microphone is converted into a numerical format compatible with the Wav2Vec2 model for transcription.

Real-time processing is a core aspect of our system, enabling immediate feedback and analysis. Facial emotion recognition occurs frame-by-frame from the live camera feed, allowing for continuous monitoring of facial expressions. On the other hand, audio emotion recognition is initiated upon the completion of audio recording, ensuring prompt analysis of spoken content.

To provide a user-friendly interface, we develop a Flask web application. This interface offers endpoints for both facial and audio emotion recognition, providing visual feedback for facial emotion recognition and interactive elements for audio emotion recognition. System deployment is achieved through the Flask web application, facilitating easy access and usage. Debug mode is enabled during development to aid in testing and troubleshooting, ensuring a smooth deployment process Integration with external services, such as the MediaPipe library for facial landmark detection and the Facebook Wav2Vec2 model for audio transcription, enhances the functionality and performance of our system. These services seamlessly complement our implementation, contributing to the overall effectiveness of emotion recognition. [14]

In our research, we integrated the "twitter-emotion-classification-with-bert" model from the Hugging Face website into our multimodal emotion recognition system for analyzing emotional states. This model is based on BERT (Bidirectional Encoder Representations from Transformers) architecture, which is well-known for its effectiveness in natural language processing tasks. We utilized this model specifically for sentiment analysis on textual data. [11]

Many researchers believe face masks are a universal application. However, other emotions, including vocalizations, show changes in characteristics such as emotional anger. Considering the importance of value, it is known that anger often has the highest value compared to the lower frequency of sadness. The difference between real and simulated data is a simple reason for this change. [4]

Figure 5. Different Facial Emotion Recognition

Our implementation is designed to support continuous improvement and adaptation. It enables iterative updates and enhancements, ensuring responsiveness to evolving requirements and user feedback.

5. Empirical Results

5.1. Performance evaluation

In evaluating the performance of our multimodal emotion recognition system, we conducted comprehensive tests to assess its accuracy, efficiency, and usability. Accuracy metrics were measured for both facial and audio emotion recognition modules separately and in combination. For facial emotion recognition, we employed standard evaluation metrics such as precision, recall, and F1-score, comparing the predicted emotions with ground truth labels. Similarly, for audio emotion recognition, we evaluated transcription accuracy and sentiment analysis performance against manually transcribed and labeled audio samples.

Furthermore, we assessed the efficiency of our system by measuring its real-time processing capabilities, including frame processing rates for facial emotion recognition and transcription speed for audio emotion recognition. These metrics were crucial in determining the system's responsiveness and suitability for real-time applications. Usability was evaluated through user feedback and Flask web application interface testing. We gathered input from users regarding the interface design, ease of use, and overall experience with the emotion recognition functionalities. Additionally, we conducted user studies to gauge the system's effectiveness in capturing and interpreting emotional states accurately.

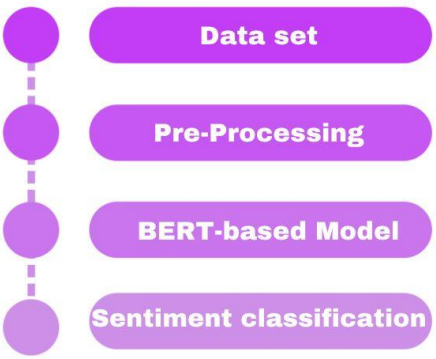

Figure 6. Emotion classification with the BERT model

Also to evaluate the performance of the "twitter-emotion-classification-with-bert" model within our multimodal emotion recognition system, we conducted rigorous testing and validation procedures. We measured various performance metrics, including accuracy, precision, recall, and F1-score, to assess the model's effectiveness in accurately classifying emotions conveyed in textual content. Additionally, we compared the model's

performance against baseline models or existing sentiment analysis approaches to ascertain its relative performance and effectiveness within our system.[11]

Overall, the performance evaluation demonstrated the effectiveness and reliability of our multimodal emotion recognition system in accurately identifying and analyzing emotional states in real-time scenarios. The results provide valuable insights into the system's capabilities and areas for potential improvement, paving the way for future advancements and applications in emotion recognition technology.

5.2. Error analysis

In assessing the multimodal emotion recognition system, we conducted a thorough error analysis to understand its performance and limitations. We identified several factors contributing to errors in emotion recognition across both facial and audio modalities. For facial emotion recognition, challenges such as varying lighting conditions, occlusions, and the complexity of facial expressions were observed, similarly, in audio emotion recognition, background noise, speech variations, and ambiguities in spoken content posed difficulties.

By examining specific instances of misclassifications, we gained insights into where the system struggled to accurately infer emotions. These included cases involving subtle facial expressions or nuanced vocal cues that the system found challenging to interpret. Identifying these areas of weakness allowed us to focus on refining preprocessing techniques and selecting more suitable models to improve performance.

Furthermore, our analysis revealed misclassification patterns across different demographic groups or emotional contexts. Certain emotions were consistently misclassified more than others, suggesting potential biases or shortcomings in the dataset or model design. To address these challenges, we explored strategies such as fine-tuning model parameters, incorporating additional data augmentation methods, and refining preprocessing steps—these efforts aimed to enhance the system's robustness and accuracy in real-world scenarios.

In the error analysis phase, we investigated the model's misclassifications and errors to gain insights into its limitations and areas for improvement. We examined specific cases where the model struggled to accurately classify emotions in tweets, identifying common patterns or challenges encountered. These errors may have resulted from ambiguous or nuanced language usage, context-dependent emotions, or limitations in the training data. By understanding these error patterns, we aimed to refine the model's architecture, fine-tune its parameters, or explore additional preprocessing techniques to enhance its performance and address specific challenges in sentiment analysis tasks. Furthermore, the error analysis provided valuable insights for optimizing the integration of the sentiment analysis model with the facial and audio modalities in our multimodal emotion recognition system, ultimately improving its overall effectiveness in real-world applications.

Overall, the error analysis provided valuable insights into the system's performance limitations and guided our efforts toward optimizing its effectiveness for emotion recognition tasks.

6. Conclusions

In conclusion, the domains of audio and speech emotion recognition stand as pivotal areas of research with transformative potential in enhancing human-computer interaction and understanding human behavior. Throughout this paper, we have explored the intricacies of various methodologies and approaches employed in these fields, ranging from traditional signal processing techniques to cutting-edge deep learning architectures. Despite remarkable progress, persistent challenges such as cross-cultural disparities in emotional expression and the scarcity of labeled data pose significant hurdles.

Further, researchers must address these challenges through interdisciplinary collaboration and innovative solutions. By developing robust models capable of generalizing across diverse cultural contexts and leveraging multimodal approaches to enhance recognition accuracy, we can advance the state-of-the-art in audio and speech emotion recognition. Additionally, integrating real-time processing capabilities is essential for practical applications such as virtual assistants, emotion-aware systems, and healthcare monitoring devices.

Acknowledgment

We express our sincere gratitude to our team members and our guide Prof. Vijesh Nair for their invaluable contributions to the development of this manuscript and the successful completion of our research project. Their

dedication, support, and expertise have been instrumental in every stage of this endeavor. We also acknowledge that this research was self-funded, and no external sponsorship was involved.

Conflict of Interest

The authors declare that there are no conflicts of interest

Author Contribution Statement

[Author1: Devesh Sali] conceptualized the project, designed the experimental framework, and contributed to the implementation of the keypoint classifier and transformer models. They also played a significant role in data preprocessing and analysis.

[Author2: Jeet Manjrekar] conducted an extensive literature review and contributed to the development of the BERT model for text-based emotion recognition. Additionally, they assisted in the integration of different modalities and the interpretation of results.

[Author 3: Sakshi Rajeshirke] focused on the audio modality, preprocessing audio data, and developing algorithms for audio-based emotion recognition. They also contributed to the experimental design and validation of the models.

[Author 4: Prathamesh Mane] contributed to the video modality by designing and implementing algorithms for extracting relevant features from video data. They also assisted in model evaluation and result interpretation across different modalities.

Sub-author (Project Guide), [Prof. Vijesh Nair], provided invaluable guidance throughout all stages of the project. They helped shape the research direction, provided critical feedback on methodology and results, and ensured the project's coherence and rigor.

Each author contributed unique expertise and effort to the project, from conceptualization to implementation and analysis. The collaboration among the authors and the guidance from the project guide was essential in the successful completion of this multi-modal human emotion recognition project.

References

[1]. Park, S. H., Bae, B. C., & Cheong, Y. G. (2020, February). Emotion recognition from text stories using an emotion embedding model. In 2020 *IEEE international conference on big data and smart computing (BigComp)* (pp. 579-583). IEEE.

[2]. Ristea, N. C., Duțu, L. C., & Radoi, A. (2019, October). Emotion recognition system from speech and visual information based on convolutional neural networks. In *2019 International Conference on Speech Technology and Human-Computer Dialogue* (SpeD) (pp. 1-6). IEEE.

[3]. Seehapoch, T., & Wongthanavasu, S. (2013, January). Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)* (pp. 86-91). IEEE.

[4]. Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE access*, 9, 47795-47814.

[5]. Sajid, M., Afzal, M., & Shoaib, M. (2021, April). Multimodal emotion recognition using deep convolution and recurrent network. In *2021 International Conference on Artificial Intelligence* (ICAI) (pp. 128-133). IEEE.

[6]. Deshmukh, R. S., Jagtap, V., & Paygude, S. (2017, June). Facial emotion recognition system through machine learning approach. In *2017 international conference on intelligent computing and control systems (iciccs*) (pp. 272-277). IEEE.

[7]. Avula, H., Ranjith, R., & Pillai, A. S. (2022, December). CNN based recognition of emotion and speech from gestures and facial expressions. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology* (pp. 1360-1365). IEEE.

[8]. Seehapoch, T., & Wongthanavasu, S. (2013, January). Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)* (pp. 86-91). IEEE.

[9]. Colaco, S., & Han, D. S. (2020, February). Facial keypoint detection with convolutional neural networks. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 671-674). IEEE.

[10]. Gavali, M. P., & Verma, A. (2023, September). Automatic Recognition of Emotions in Speech With Large Self-Supervised Learning Transformer Models. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)* (pp. 1-7). IEEE.

[11]. Zhang, T., & Zhang, R. (2021, November). Revealing the power of BERT for text sentiment classification. In *2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)* (pp. 14-17). IEEE.

[12]. Pattun, G., & Kumar, P. (2023, December). Emotion Classification using Generative Pre-trained Embedding and Machine Learning. In *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)* (pp. 1-6). IEEE.

[13]. Kwon, J., Oh, K. T., Kim, J., Kwon, O., Kang, H. C., & Yoo, S. K. (2023, December). Facial Emotion Recognition using Landmark coordinate features. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 4916-4918). IEEE.

[14]. Soky, K., Li, S., Chu, C., & Kawahara, T. (2023, June). Domain and language adaptation using heterogeneous datasets for wav2vec2. 0-based speech recognition of low-resource language. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
[15]. Ghofrani, A., Toroghi, R. M., & Ghanbari, S. (2019, February). Realtime face-detection and emotion recognition using mtcnn and minishufflenet v2. In *2019 5th conference on knowledge based engineering and innovation* (KBEI) (pp. 817-821). IEEE.
[16]. Avabratha, V. V., Rana, S., Narayan, S., Raju, S. Y., & Sahana, S. (2024, July). Speech and Facial Emotion Recognition using Convolutional Neural Network and Random Forest: A Multimodal Analysis. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1-5). IEEE.
[17]. Qin, G., Zhu, Y., Wu, Z., Jiang, Q., Yin, J., Sun, J., ... & Wang, Y. (2024, May). Application of Convolutional Neural Network in Multimodal Emotion Recognition. In *2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT)* (pp. 440-444). IEEE.
[18]. Sun, D., He, Y., & Han, J. (2023, June). Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
[19]. Deshmukh, S., Chaudhary, S., Gayakwad, M., Kadam, K., More, N. S., & Bhosale, A. (2024, April). Advances in Facial Emotion Recognition: Deep Learning Approaches and Future Prospects. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)* (pp. 1-3). IEEE.
[20]. Liu, Y., Geng, D., Wu, X., & Liu, Y. (2024, September). Multimodal Emotion Recognition based on Convolutional Neural Networks and Long Short-Term Memory Networks. In *2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC)* (pp. 69-73). IEEE.
[21]. Zaidi, S. A. M., Latif, S., & Qadir, J. (2024). Enhancing Cross-Language Multimodal Emotion Recognition With Dual Attention Transformers. *IEEE Open Journal of the Computer Society*.
[22]. Yadav, U., Bondre, S., Thakre, B., & Likhar, K. (2024, July). Speech-to-text Emotion Detection System using SVM, CNN, and BERT. In *2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)* (pp. 1-5). IEEE.
[23]. Rana, S., Chaudhary, R., Gupta, M., & Garg, P. (2023, December). Exploring Different Techniques for Emotion Detection Through Face Recognition. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 779-786). IEEE.
[24]. Betageri, D., & Yelamali, V. (2024, July). Detection and Classification of Human Emotion Using Deep Learning Model. In *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)* (pp. 1-5). IEEE.
[25]. Negi, A. S., Arora, A., Bisht, S., Devliyal, S., Kumar, B. V., & Kaur, G. (2024, April). Facial Emotion Detection using CNN & VGG16 Model. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1-6). IEEE.
[26]. Avabratha, V. V., Rana, S., Narayan, S., Raju, S. Y., & Sahana, S. (2024, July). Speech and Facial Emotion Recognition using Convolutional Neural Network and Random Forest: A Multimodal Analysis. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1-5). IEEE.
[27]. Sharma, A., Bajaj, V., & Arora, J. (2023, February). Machine learning techniques for real-time emotion detection from facial expressions. In *2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (pp. 1-6). IEEE.
[28]. Ruangdit, T., Sungkhin, T., Phenglong, W., & Phaisangittisagul, E. (2023, October). Integration of Facial and Speech Expressions for Multimodal Emotional Recognition. In *TENCON 2023-2023 IEEE Region 10 Conference (TENCON)* (pp. 519-523). IEEE.
[29]. Basavaiah, J., Anthony, A. A., HN, N. K., & Patil, C. M. (2024, March). Facial Emotion Recognition: A Review on State-of-the-art Techniques. In *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)* (pp. 1-6). IEEE.
[30]. Cioroiu, G., & Radoi, A. (2024, October). Emotion Recognition from Contextualized Speech Representations using Fine-tuned Transformers. In *2024 15th International Conference on Communications (COMM)* (pp. 1-5). IEEE.
[31]. Abakarim, F., & Abenaou, A. (2024, May). Speech Emotion Recognition System Using Discrete Wavelet Transform and Support Vector Machine. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-5). IEEE.
[32]. Alqurashi, N., Li, Y., & Sidorov, K. (2024, June). Improving speech emotion recognition through hierarchical classification and text integration for enhanced emotional analysis and contextual understanding. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
[33]. Konappanavar, T. S., Loni, J. S., Adhyapak, S., & Patil, S. B. (2023, November). Real-Time Facial Emotion Detection Using Machine Learning. In *2023 2nd International Conference on Futuristic Technologies (INCOFT)* (pp. 1-5). IEEE.
[34]. Liu, H., Huang, X., & Xu, N. (2024, May). Research on Emotion Recognition Model Construction and Emotion Regulation Technology Based on Machine Learning Algorithm. In *2024 International Conference on Telecommunications and Power Electronics (TELEPE)* (pp. 850-854). IEEE.
[35]. Bakkialakshmi, V. S., Kar, N., & Kumar, V. (2024, March). The Digital Mirror-Reflecting Human Emotions through Machine Learning-Based Facial Gesture Recognition. In *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)* (pp. 01-06). IEEE.
[36]. https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html