Generative AI-Based Multimodal Sentiment Analysis of Low-Resource Languages

Mohammad Usman Zafar[1], Hamid Ghous[2], Sana Jamshaid[1], Mubasher H. Malik[1]

[1]Vision, Linguistics and Machine Intelligence Research Lab, Multan, Pakistan
[2]Australian Scientific and Engineering Solutions, Australia

Abstract: Multimodal sentiment analysis (MSA), which integrates emotional computing approaches across textual, visual, and audio modalities, has become a crucial tool in understanding human behavior. The majority of MSA research, however, is on high-resource languages, which leaves a big gap in the investigation of attitudes among groups speaking low-resource languages like Saraiki and Punjabi. This study addresses a significant problem in the field of sentiment analysis for low-resource languages, which frequently lack enough annotated datasets and strong tools compared to high-resource languages. By proposing a Generative AI-driven framework that trades on fake data generation, transfer learning, and multimodal fusion techniques, the research offers a transformative approach to overcoming these restrictions. This study introduces a generative AI-based multimodal framework for sentiment analysis of low-resource languages, with a focus on Punjabi and Saraiki. Manually created two video datasets. Using these videos, both audio and representative frames were extracted. To address the dataset shortage, Generative Adversarial Networks (GANs) were applied to generate synthetic audio and frames, thereby improving data range and class balance. The process of feature extraction was done using Wav2Vec2 for audio and ResNet-50 for frames. The extracted embeddings were further categorized using both traditional machine learning methods and deep learning approaches. On the whole, the experimental findings indicated that deep learning classifiers performed better in comparison to traditional machine learning models in both Punjabi and Saraiki datasets. CNN and BiLSTM had the best F1-scores of 0.66 and 0.47, and the respective AUC of 0.83 and 0.67, which indicates the model is very good at multimodal sentiment analysis. The research emphasizes the efficiency of multimodal and generative approaches for sentiment analysis in underrepresented languages and suggests directions for future research.

Keywords: Machine Learning, Deep Learning, Punjabi, Saraiki, Generative AI

Email: mubasher@usp.edu.pk

1. Introduction

Even though there has been improvement in sentiment analysis for high-resource languages, low-resource languages stay underrepresented owing to challenges such as a shortage of, limited linguistic resources, labeled datasets, as well as a shortage of tools capable of addressing their unique syntactic and semantic (Ali, 2021) [1]. Current Generative AI models and multimodal sentiment analysis techniques are primarily optimized for high-resource settings, leaving low-resource languages underserved. The combination of multimodal data, combining text, visual, and auditory inputs, presents an opportunity to improve sentiment prediction. However, the lack of effective frameworks that force Generative AI and multimodal learning for low-resource languages limits the ability to bridge this space (Biswas, 2024) [2]. This study seeks to address these challenges by introducing a Generative AI-based

multimodal framework modified for low-resource languages, focusing on reducing reliance on extensive annotated datasets, leveraging transfer learning, and combining cross-modal awareness mechanisms to increase sentiment analysis performance and inclusivity (Lupascu, 2025) [3]. Subdivision of artificial intelligence (AI) is Natural Language Processing and CS that employs ML to make it feasible for computers to understand and also verbalize human language (Ali, 2021) (Rongali, 2025) [4]. Statistical modeling and ML, DL, and a combination of computational linguistics, as the rule-based modeling of human language, enables Pc and digital devices to detect, comprehend, and produce speech (Alshahrani, 2022) [5]. As of the communication capability of (LLMs) to the facility of image-making models to understand requests, Natural language processing research has contributed to the emergence of Generative AI (Carvalho, 2019) [6]. Several people presently use NLP in their everyday routine, the same as observed by search engines, voice-activated chatbots in favor of consumer help, sound-activated global positioning systems, and smartphone digital assistants that can reply to questions, like Siri, and Apple's (Kumar, 2023) [7]. In enterprise explanations that assist in mechanizing and streamlining business operations, boosting worker productivity, and making business procedures simpler, NLP is becoming more important. (Stryker, 2024) [8] Multimodal AI is the word used to illustrate machine learning models that are able to process and integrate data from several modalities or types (Kumar V. , 2022) [9]. Text, images, sound, video, and other sensory inputs are examples of these modalities. Compared to normal AI models, which are normally designed to handle a particular type of data, multimodal AI integrates and analyzes several data inputs to create more detailed comprehension and strong outputs. Multimodal AI systems are able to withstand missing data and noise better. The system can rely on other modalities to continue operating even if one is unavailable or unreliable. By facilitating more intuitive and natural interfaces for improved user experiences, multimodal AI improves human-computer interaction. For example, spoken orders and visual cues can both be understood and responded to by virtual assistants, which facilitates more efficient and seamless interactions. Consider a bird identification software that can identify pictures of a certain bird and verify its identity by "listening" to a sample of its song, or a chatbot that can discuss your spectacles and suggest a size based on a picture you send it. Users can interact with data in more ways and receive more relevant outputs from AI that can function across several sensory dimensions. (Stryker, ibm, 2024) [10] The term "Generative AI" explains deep learning models that, given the data they were trained on, are able to create exceptional text, photos, and other content. Even among doubters, the introduction of ChatGPT appears to be a watershed in the history of artificial intelligence,

despite its repeated rounds of hype. With the help of its most modern wide language model, OpenAI's chatbot is able to produce essays, poems, and jokes that pass for human-written content. With just a few phrases, prompt ChatGPT to produce Nick Cave-style song lyrics or Yelp reviews that are love poetry (Goodlad, 2024) [11]. AI that can make creative material such as text, photos, audio, videos, or appeal or program source code in reaction to a user's request is known as Gen AI. Gen AI is based on DL model algorithms that remove learning and management processes in the human brain. To process natural language queries or questions made by users and provide relevant new material in response, these models initially locate and encode relationships and patterns in big size of data. (Scapicchio, 2024) [12]. DL is a form of ML that simulates the challenging process of making decisions by the human brain through multifaceted neural networks (Cai, 2021) [13]. The collection of AI apps in our everyday lives is power-driven by DL in one method or another (Gupta, 2023) [14]. The topology of the original neural network architecture is the main distinction between ML and DL (Ahmed, 2023) [15]. In "nondeep," basic neural networks with one or two computational layers are employed. conventional ML models. DL models are trained on 3 or more layers, but normally 100 or 1000. Supervised learning models need structured and labeled input data to give consistent output, and Unsupervised learning is possible with deep learning models (Alloghani, 2020) [16]. Unsupervised learning can be applied in DL models to derive the traits, attributes, and relationships that may produce accurate outputs of unstructured, raw data (James, 2023) [17]. For better precision, these models have the ability to evaluate and improve their outputs. Deep learning is an aspect of data science that drives several services and applications that are more likely to further automation through the conduct of physical and analytical tasks without human involvement (Jabeen, 2023) [18]. Voice-activated Digital assistants, credit card fraud detection, TV remote controls, and Gen AI are just a small number of the commonplace goods and services made possible by this. (Holdsworth, 2024) [19] For understanding human emotions in text, images, and speech, sentiment analysis has become a fundamental tool. Nevertheless, a challenge faced by low-resource languages is limited datasets and tools (Guo, 2022) [20]. This research proposes a new approach leveraging Generative AI for multimodal sentiment analysis in low-resource languages. By making use of transfer learning, artificial data augmentation, and multimodal fusion techniques, the framework achieves significant improvements in accuracy and generalization ability (Fayaz, 2024) [21]. Experiments on custom datasets demonstrate the potential of the approach to bridge the gap in sentiment analysis for underrepresented languages

2. Literature Review

2.1 Multimodal Sentiment Analysis

Han (Han, 2021) [22] recommends the Bi-Bimodal Fusion Network (BBFN), a novel end-to-end network exploiting pair-wise representation of modality to achieve separation (difference increment) and fusion (relevance increment). The main focus of that field of research is creating a remarkable fusion system that can combine and extract important data from multiple modalities.

In multimodal learning, representation learning is an important and difficult issue (Yu, 2021) [23]. To learn independent unimodal supervisions, create a label generation module based on the self-supervised learning approach. Then, in order to learn consistency and difference, respectively, the unimodal and multimodal tasks are jointly trained. The author provides a detailed explanation of the Self-Supervised Multi-task Multimodal sentiment analysis network. Alderazi (Alderazi, 2024) [24] creates strategies to address the deficiency of topic-based labeling methods, evaluates several methods for creating big, annotated datasets, and assesses how well large language models (LLMs), deep learning (DL), and machine learning (ML) perform in classifying Arabic textual data. The author uses Machine learning, Naive Bayes (NB), SVM, KNN, Deep Learning Model, Generative AI (ChatGPT), and Balancing Data. For the problem of sentiment and emotion-controlled dialogue production, (Firdaus, 2020) [25] first presents a large-scale benchmark Sentiment Emotion aware Multimodal Dialogue (SEMD) dataset. This paper (Krugmann, 2024) [26] provides a thorough examination of LLMs' competence in sentiment analysis, a crucial research assignment in marketing that helps identify the feelings, thoughts, and impressions of consumers. Three cutting-edge LLMs, GPT-3.5, GPT-4, and Llama 2, are evaluated against well-known, highly effective transfer learning models. (Yadav, 2023) [27] introduces DMLANet, a network designed to enhance multimodal sentiment analysis by focusing on complex correlations between image and text data. DMLANet uses a multi-level attention mechanism to improve feature extraction, combining channel attention and spatial attention in visual data to form a "bi-attentive" feature map. Hossain (Hossain, 2022) [28] introduces MemoSen, a new multimodal dataset for Bengali that includes 4368 memes with the sentiment labels "positive," "negative," and "neutral" annotated. MemoSen is used to conduct a series of studies by building 10 multimodal (image+text) and twelve unimodal (visual, textual) models, such as MemoSen, which is the name of the Sentiment Labels, and Bengali is the language (including code-mixed and code-switched texts, or Banglish). 4,368 memes in size. Sources Face book, Twitter, and Instagram.

Ghandi (Gandhi, 2023) [29] provides an in-depth review of the evolution and advancements in multimodal sentiment analysis (MSA). It explores how combining modalities such as text, audio, and visual data enhances sentiment detection accuracy, leveraging cutting-edge machine learning and deep learning techniques.

Thakkar (Thakkar, 2024) [30] fills this gap by using a simple curation procedure to convert an existing textual Twitter sentiment dataset into a multimodal version. Our approach creates new opportunities for the research community to study sentiment. Furthermore, the author uses this enhanced dataset to perform baseline experiments and reports the results. Text encoders utilize nearly fifty data sets. By utilizing a late fusion technique to combine the three distinct modalities for the initial sentiment prediction, (Das, 2023) [31] suggests the multimodal sentiment analysis framework, also known as Textual Visual Multimodal Fusion (TVMF). The Gujarati Sentiment Analysis Corpus (GSAC), which was gathered from Twitter and manually annotated by native speakers of the language, is presented and described by (Gokani, 2023) [32]. In order to provide trustworthy baselines for future work, the author conducts extensive experiments on his corpus and describes in detail his collection and annotation processes.

2.2 Deep Learning

Kamal (Kamal, 2024) [33] uses the CMU Multimodal Opinion Sentiment Expression in Recordings CMU and the CMU MOSEI Multimodal Sentiment Opinions (CMU MOSI) dataset benchmark to evaluate 8 DL based multimodal models for sentiment analysis. Beyond illustrating the potential of multimodal sentiment analysis, the author provides insightful information. Jabeen (Jabeen, 2023) [34] emphasizes a variety of modalities, including text, audio, video, images, body language, facial expressions, and physiological cues. For a review, the author uses a variety of datasets, including Multimodal Image Description, Multimodal Video Description, and Multimodal Emotion Recognition (MMER) for applications and techniques in Multimodal DL Jabeen suggests fusion techniques, model architectures, deep reinforcement learning, and multimodal deep learning techniques. Liu (Liu, 2024) [35] investigates the use of NLP tools based on DL in multilingual sentiment analysis. Multilingual sentiment datasets encompassing key languages like Spanish, Chinese, and English are used in the study. Research shows how well deep learning models like BERT and LSTM perform in cross-linguistic sentiment categorization tasks when presented in multilingual environments. Ahmad (Ahmed, 2024) [36] offers DL techniques for sentence-level Urdu sentiment analysis for MIoT. The approach of the author consists of a variety of phases, i.e., ext preprocessing, data gathering, testing, model training, and evaluation. Most of the data

sets are built for lexicon-based Urdu SA approaches; the IMDB standard data set and the Urdu blogs data set are used for a new approach for Sentiment Analysis of a low-resource language using DL models. In order to tackle issues such as data source heterogeneity, information fusion, modality alignment, and synchronization, and the development of efficient extract feature techniques that take into custody discriminative information from all modalities, (Aslam, 2023) [37] presents a new approach titled attention-based Multimodal Sentiment Analysis and Emotion Recognition. Bashir (Bashir, 2023) [38] Significant efforts have been made to identify emotions in textual data in Chinese that is high resource language, French, English, and others. The dataset consists of sentences and paragraphs annotated with various emotions, and is made publicly available in this work, The Urdu Nastalique Emotions Dataset. Sharma (Sharma, 2018) [39] provides a comparative examination of multimodal sentiment analysis in the company of a network for deep neural that incorporates NLP and image recognition. The author gathers MOSI, Twitter image, and movie review datasets for deep learning-based multimodal sentiment analysis. Abdu (Abdu, 2021) [40] provides a thorough classification of 35 cutting-edge models that have recently been put forth in the field of sentiment analysis for video into 8 groups according to the architectures that each model uses. (Xu, 2017) [41] For multimodal sentiment analysis, suggest MultiSentiNet, a deep semantic network. The MultiSentiNet model's efficacy is confirmed by experiments conducted on two publicly accessible sentiment datasets, MVSA (Multi-View Sentiment Analysis)-Single: 5229 image-text pairs and MVSA-Multi 18,600 text-image pairs. Khan (Khan, 2021) **[**42] creates a dataset benchmark similar to Urdu Corpus for Sentiment Analysis (UCSA), which consists of 9,601 reviews from sources such as dramas, politics, movies, sports, TV discussion shows, and software for sentiment analysis in a resource-poor language like Urdu. The author assesses different DL and machine learning methods for sentiment. In order to conduct experiments for all feature types, (Khan L. A., 2021) [43] takes into consideration a set of DL classifiers (1D-CNN and LSTM) and machine learning classifiers (SVM, RF, AdaBoost, NB, LR, and MLP).

2.3 Machine Learning

(Malviya, 2020) [44] The TEQIP Collaborative Research Scheme, "Develop an Efficient Machine Learning-based Approach for Sentiment Analysis using Online Videos," has its heartfelt gratitude. For sentiment analysis, the authors employed deep learning techniques like CNN, RNN, LSTM, and BI-LSTM in addition to traditional ML models like Support Vector Machines (SVM) and Random Forest. Faria (Faria, 2024) [45] uses a variety of features and both deep learning and classical models, such as Gaussian naive Bayes (GNB),

random forests (RFs), support vector machines (SVMs), multilayer perceptrons (MLPs), and a 1D convolutional neural network, to precisely identify and classify emotions in speech. An overview of the techniques used to conduct sentiment analysis across languages is provided by (Mercha, 2023) [46]. By utilizing various datasets, such as rich-resource languages, low-resource languages, multilingual, and cross-lingual approaches. Khan (Khan, 2024) [47] provides a thorough investigation of sentiment analysis in Urdu with an emphasis on a machine learning model designed especially for Urdu text. For Resource Deprived Language Sentiment Analysis Employing Machine Learning and Feature Extraction Techniques, An Analysis of Urdu Writers' Use of IMDB Reviews. 50,000 movies are included in the Urdu Text Sentiment Analysis Dataset (UTSA) and the Urdu Translated Dataset (IMDb_Urdu).

2.4 Research Gap

Based on the review of literature related to the topic in Chapter 2, it is a very challenging task that is sentiment analysis of low-resource languages like Punjabi and Saraiki, especially because of the lack of annotated data and the underutilization of multimodal approaches. These research gaps form the foundation of this study and motivate the need for developing effective methodologies. Therefore, Chapter 3 focuses on the proposed techniques, detailing the preprocessing steps, methods of feature extraction, and DL and ML models employed to address the identified challenges and advance sentiment analysis in low-resource contexts.

3. Methodology

After the literature review, the methodology is discussed in this chapter. Research Techniques is the process of finding, analyzing information, and selecting data about the topic. The methodology of this research allows the person who reads to judge the study’s overall validity and dependability.

The methodology of this study aims at formulating and testing a Generative AI-based multimodal sentiment analysis model of low-resource languages, particularly Punjabi and Saraiki. The research is based on a quantitative research design because the main source of data is the hand-made datasets of videos. All videos belong to the three classes of sentiment, and these are positive, negative, and neutral. The methodology includes several steps, the first of which is data preprocessing, where the audio, visual, and textual data are obtained from the videos. Each modality is then subjected to feature extraction: extracted audio features with the help of models such as Wav2Vec2, and visual ones are extracted with the help of CNNs such as ResNet50. Lastly, sentiment classifiers are done with ML and DL classifiers. Standard measures of model performance are F1-score, recall, accuracy, and AUC-ROC curves. Multimodal and generative approaches ensure that this systematic strategy will be able to cope

with the specific challenges of improving accuracy, generalizability, and sentiment analysis in low-resource languages.

Methodology contains the following steps:

- Datasets
- Importance of data pre-processing
- Methods
- Proposed Framework

3.1 Dataset:

In this study, two different datasets could be generated, which included video data in the Punjabi and Saraiki languages. We manually developed each dataset so that they would be culturally and linguistically relevant, among 3 classes of sentiment, namely neutral, positive, and negative. The total amount of the Punjabi dataset is 310 videos, comprised of 87 positive samples, 77 negative ones, and 146 neutral samples. Likewise, the Saraiki data database has 155 videos, with the distribution 60 positive, 43 negative, and 52 neutral. We take two video-based multilingual datasets, Punjabi and Saraiki. An audio track was obtained in both videos to record speech-based sentiment cues, and 20 representative frames were taken in each video to save visual data to be used in later analysis. Both datasets were constructed with great care in order to capture genuine manifestations of sentiment in contexts, and hence they can be used in multimodal sentiment analysis. The dataset was validated by a field expert. Being created by the researcher, such resources provide a new and valuable piece of data concerning the study of low-resource languages.

Table 1: Punjabi dataset for video, audio, and frames

Punjabi Videos	**No of Videos**	**Punjabi Audios**	**No of audios**	**Punjabi Frames**	**No of Frames**
Punjabi Positive	87	Punjabi Positive	87	Punjabi Positive	1740
Punjabi Neutral	146	Punjabi Neutral	146	Punjabi Neutral	2920
Punjabi Negative	77	Punjabi Negative	77	Punjabi Negative	1540

Table 2: Saraiki dataset for video, audio, and frames

Saraiki Videos	**No of Videos**	**Saraiki Audios**	**No of audios**	**Saraiki Frames**	**No of Frames**
Punjabi Positive	60	Punjabi Positive	60	Punjabi Positive	880
Punjabi Neutral	52	Punjabi Neutral	52	Punjabi Neutral	640
Punjabi Negative	43	Punjabi Negative	43	Punjabi Negative	580

3.2 Preprocessing

The preprocessing stage was associated with gathering audio and visual content of the raw video files of Saraiki and Punjabi languages to prepare them to be analyzed multimodally regarding sentiment. To maintain the sentiment indicators in speech to be extracted later on, the MoviePy library (moviepy.editor) was applied to extract the audio component of each video. Simultaneously, the visual data was prepared by extracting 20 typical frames of each video with the help of the OpenCV (cv2) library. The selection of these frames was to give a good background to the visual sentiment analysis, as many expressions of faces and contextual information were captured in the movie. This preprocessing step was important in order to convert the raw video information into structured audio and image input, in order to extract the features and integrate multimodal features effectively in further stages of the research.

3.2.1 Audio Extraction

Each Punjabi and Saraiki video was processed with the MoviePy library (moviepy.editor) in order to extract audio tracks. This step separated the speech signals of the raw video data, allowing features that were related to sentiment, e.g., tone, pitch, and prosody, to be recorded in later analysis. $min\ max\ V\ (D, G) = \mathrm{E}_{x\sim p\ data^{(x)}}[\log D\ (x)] + Ez\sim p_{z^{(z)}}\left[\log(1 - D\big(G(z)\big))\right]$

3.2.2 Frame Extraction

20 representative frames were selected from each video and converted to visual information with the OpenCV (cv2) library in order to save them. These frames were selected to capture diverse expressions, gestures, and contextual cues that contribute to visual sentiment recognition. The extracted frames provide a static but informative representation of the video data, facilitating further feature extraction through deep learning models. $y(i,j) = \sum_m \sum_n x(i+m, j+n).w(m,n) + b$

3.3 Methods

Methods are clearly defined processes used to accomplish a given purpose, and they can have distinct meanings as well. For Examples of specific algorithms, approaches, or models used to evaluate data, define or predict outcomes, and improve performance. The proposed study combines feature extraction, classification, and generative artificial intelligence to create a strong framework for multimodal sentiment analysis in Saraiki and Punjabi. To tackle the problem of limited data in low-resource languages, more synthetic data was produced in both modalities, that is, audio and video frames, using a Generative Adversarial Network (GAN) technique. By improving the dataset's size and balance, this augmentation step improved its suitability for deep learning model training.

By using Convolutional Neural Networks (CNNs), specifically the ResNet-50 architecture, features were extracted from video frames for the visual modality, which is known for its ability to capture discriminative spatial patterns from images. For the audio modality, high-level representations were obtained using Wav2Vec2, a transformer-based deep learning model designed for speech processing, which effectively captures phonetic, prosodic, and linguistic cues from raw audio signals. Once features were extracted, sentiment classification was performed by a range of both ML and DL methods to ensure comprehensive evaluation. (Li, 2021) [48]

As baseline classifiers, ML techniques, including Decision Tree, Support Vector Machine, and Random Forest, were implemented. In parallel, DL models, including LSTM, Bi-LSTM, Transformer, and CNN-based architectures, were applied to leverage temporal, sequential, and contextual dependencies within the multimodal data. A hybrid methodological strategy provides the framework to capture complementary strengths of different models, ensuring improved performance and reliability in classifying sentiments into positive, negative, and neutral categories. (Qaiser, 2021) [49]

3.3.1 GAN

GANs are a type of generative technique that may create novel data samples that are similar to an existing dataset. To produce realistic synthetic data, GANs seek to simulate the underlying data distribution, in contrast to conventional ML methods that concentrate on classification or prediction. For low-resource languages like Punjabi and Saraiki, this makes GANs extremely helpful in tasks like speech synthesis, image production, and data augmentation. A GAN: a discriminator (D) and a generator (G) make up two neural networks.

Generative Adversarial Networks have found legitimate applications in several areas, such as computer vision, speech processing, and NLP. GANs can be applied to create artificial audio or video samples, which can be used to expand sentiment analysis datasets, especially in low-resource languages. GANs can enhance the performance of downstream classifiers like CNNs, LSTMs, and Transformers by enriching the dataset. Because of this, GANs are a crucial instrument for resolving issues with data scarcity in multilingual and multimodal sentiment analysis studies. (Goodfellow, 2020) [50]

3.3.2 WaveGAN

A particular kind of Generative Adversarial Network (GAN) called WaveGAN is made to directly generate raw audio waveforms without the need for spectrogram conversion. Chris Donahue first used it to create audio samples of speech, music, and ambient noises. WaveGAN maps a random noise vector into a realistic audio waveform using 1D transposed convolutions,

also known as deconvolutions, in contrast to image GANs, which employ 2D convolutions. This guarantees that realistic temporal structures in audio are learned by the generator. (Reddy, 2023) [51]. $z = f(x)$ where f is a CNN-based feature encoder.

3.3.3 Res Net-50

DL models known as CNNs are mainly good at tasks involving images, video frames, and other spatially structured data. To learn hierarchical feature representations, the convolutional, pooling, and fully connected layers of a CNN are used. ResNet-50, a deep residual network with 50 layers, is one of the most significant CNN architectures according to He et al. By adding residual connections, ResNet-50 solved the vanishing gradient issue that arises in very deep neural networks and improved the flow of gradients during back propagation.

$$y(i,j) = \sum_{m} \sum_{n} x(i+m, j+n).w(m,n) + b$$

3.3.4 MFCC

In speech and audio processing, Mel-Frequency Cepstral Coefficients (MFCCs) are often utilized features, particularly for tasks like sentiment analysis, speaker identification, and voice recognition. Based on how the human ear perceives frequencies, MFCCs depict the short-term power spectrum of sound. Since the Mel scale is more in accordance with how people hear pitch, MFCCs use it rather than the linear frequency scale. MFCCs ignore extraneous information like background noise and reduce dimensionality to capture the most significant aspects of speech. For audio-based machine learning applications, such as sentiment analysis in low-resource languages like Saraiki and Punjabi, they are therefore perfect.

$$F(w) = \int -\infty\infty f(t)e - i\omega t dt$$

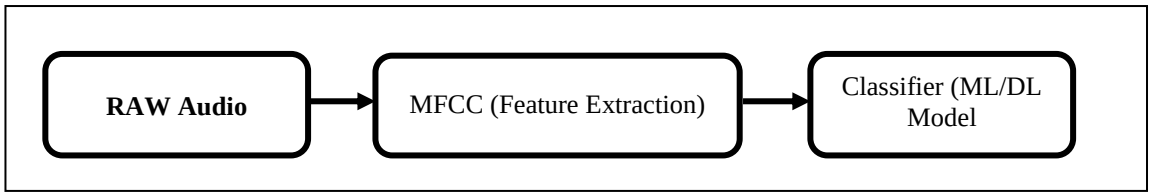

Figure: 1

3.3.5 Wav2Vec2

A deep learning model called Wav2Vec2 was created by Facebook AI (Meta) for self-supervised representation learning using unprocessed audio waveforms. Because Wav2Vec2 learns directly from raw speech signals rather than using artificial characteristics like MFCCs, it is far more effective for low-resource languages like Saraiki and Punjabi. The model is composed of a

Transformer-based context network that records long-range dependencies in the audio sequence after a multi-layer convolutional encoder converts raw waveforms into latent representations. Wav2Vec2 is therefore perfect for extracting significant speech features for use in subsequent tasks such as emotion detection, speech.

$$\mathscr{L} = -log\frac{\exp(sim(c, z^{+})/k}{\sum_{z-\in N}\exp(\mathrm{sim(c,z-)/k})}$$

3.4 Machine Learning Classifiers for Sentiment Classification

ML classifiers such as SVM, Decision Trees, and Random Forests.

3.4.1 Decision Tree (DT)

In supervised machine learning, decision trees are among the most popular algorithms because of their ease of use, interpretability, and capacity to handle both continuous and categorical data. A DT is a tree format, which is a representation of rules obtained based on the data. Each node is a feature-based decision, and each branch is the outcome of the result. In sentiment analysis, Decision Trees are often used as baseline classifiers because they provide an interpretable model that shows exactly which features drive classification decisions. For your research on Punjabi and Saraiki, Decision Trees can help highlight which aspects of speech embeddings (such as frequency bands or learned contextual features) contribute most to identifying sentiment polarity. However, one limitation of Decision Trees is that they tend to overfit training data when the tree grows too deep. This is why ensemble methods like Random Forest are often preferred in large-scale or noisy datasets.

- Entropy: $H(S) = -\sum_{i=1}^{n} pi \log_2 (pi)$ (Equation 7)
- Information Gain: $IG(s, A) = H(S) - \sum_{v\in Values(A)} \frac{|S_v|}{|S|} H(S_V)$

3.4.2 Random Forest (RF)

Random Forest is called an ensemble technique that builds on DT to increase the robustness and performance of predictions. As an alternative to relying on a single tree, Random Forest combines the predictions of many DT to achieve a consensus result. This ensemble approach reduces overfitting and variance, making Random Forest a highly reliable algorithm for classification tasks. It is particularly powerful for tasks where the dataset is diverse and noisy, as is often the case with real-world audio sentiment data in low-resource languages like Punjabi and Saraiki. In the context of sentiment classification, for example, certain trees may prefer linguistic embeddings from Wav2Vec2 while others may emphasize prosodic parameters (pitch, energy). The combination of these trees' judgments results in a sentiment label that is more accurate.

(Brennan, 2024) [52] Random Forest is a strong classifier in my study that can manage the unpredictability of Saraiki and Punjabi audio data. The flexibility of Random Forest to average over numerous models aids in producing reliable findings because the datasets for these languages are limited and prone to class imbalance. Random Forest prediction (classification): $\acute{y} = mode\{h_1(\mathrm{x}),h_2(\mathrm{x}),\ldots,h_m(\mathrm{x})\}$

where $h_1(x) is\ the\ prediction\ from\ the\ i^{th}$ decision tree

3.4.3 SVM

Support Vector Machine (SVM) is a powerful supervised learning algorithm that has proven effective in classification problems involving high-dimensional data. The main idea behind SVM is to find the optimal hyperplane that separates classes with the maximum possible margin, thereby minimizing classification error. Support vectors are the data points that are closest to the decision boundary and are identified by the SVM classifier. For linearly separable data, a linear hyperplane is sufficient. Linear, polynomial, and radial basis function (RBF) kernels are frequently used. In an abstract feature space, for instance, an RBF kernel can distinguish between Positive, Neutral, and Negative embeddings even if they overlap in the original space. (Ibm, 2023) [53] SVM is especially well-suited for my Punjabi and Saraiki sentiment classification problem since it effectively manages high-dimensional embeddings and frequently performs well even with sparse training data. SVMs may use pre-extracted embeddings (from Wav2Vec2 or ResNet-50) and get competitive results, whereas deep learning models need very large datasets to obtain significant generalization.

- Liner SVM optimization: $\min_{w,b} {}^{1}/_{2}\,||w||^2 \quad s.t.\, y_i\,(w.x_i+b) \geq 1 \forall i$
- Decision function: $f(x) = \mathrm{sign}(w.x + b)$

3.5 Deep Learning Classifiers for Sentiment Classification

Deep learning model like Deep learning models, such as CNN, LSTM, Bi-LSTM, and Transformers, automatically learn complex features from audio and video data. They capture contextual and temporal information better than traditional methods and are well-suited for multimodal sentiment analysis, especially in low-resource languages.

3.5.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are popular deep learning models that function well with sequence and auditory data. They were first created for picture recognition. They can extract spatial or temporal information from input signals since they are founded on the ideas of weight sharing and local connectedness. CNNs can identify discriminative patterns in

spectrograms, MFCCs, or feature embeddings like Wav2Vec2 outputs for sentiment analysis. CNNs use learnable filters to apply convolution operations to the input, then pooling layers and nonlinear activation functions (ReLU) to minimize dimensionality. In sentiment detection tasks, they offer robustness and computational efficiency. However, recurrent models like LSTM are better at modeling long-range dependencies than CNNs. (Zoumana, 2013) [54]

Convolution Operation

$$h_{i,j} = \sum_{m=o}^{M-1} \sum_{n=0}^{N-1} x_{i+m.j+n} \cdot w_{m,n} + b$$

3.5.2 Long Short-Term Memory (LSTM)

Recurrent neural networks (RNNs) of the LSTM network type are made to model sequential data while resolving the vanishing gradient issue. Their ability to capture long-term dependencies within sequential embeddings, like MFCC features or Wav2Vec2 outputs, makes them ideal for sentiment analysis of both text and audio.

3.5.3 Bidirectional LSTM (Bi-LSTM)

An extension of LSTM that can process input sequences both forward and backward is called bidirectional LSTM (Bi-LSTM). Because of this, the model is able to capture both past and future context, which makes it especially useful for jobs where classification relies on complete sequence information. Two parallel LSTMs make up bi-LSTMs; one processes the input from left to right, while the other does the same from right to left. Bi-LSTMs are useful for sentiment analysis in Punjabi and Saraiki audio because they capture both the evolution of sentiment over time and the ways in which future context affects how previous speech is interpreted. Comparing this to conventional LSTMs, the classification is more accurate. (sciencedirect) [55]

Bi-LSTM Output

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_t} - 1)$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_t} + 1)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

Proposed Feature Diagram:

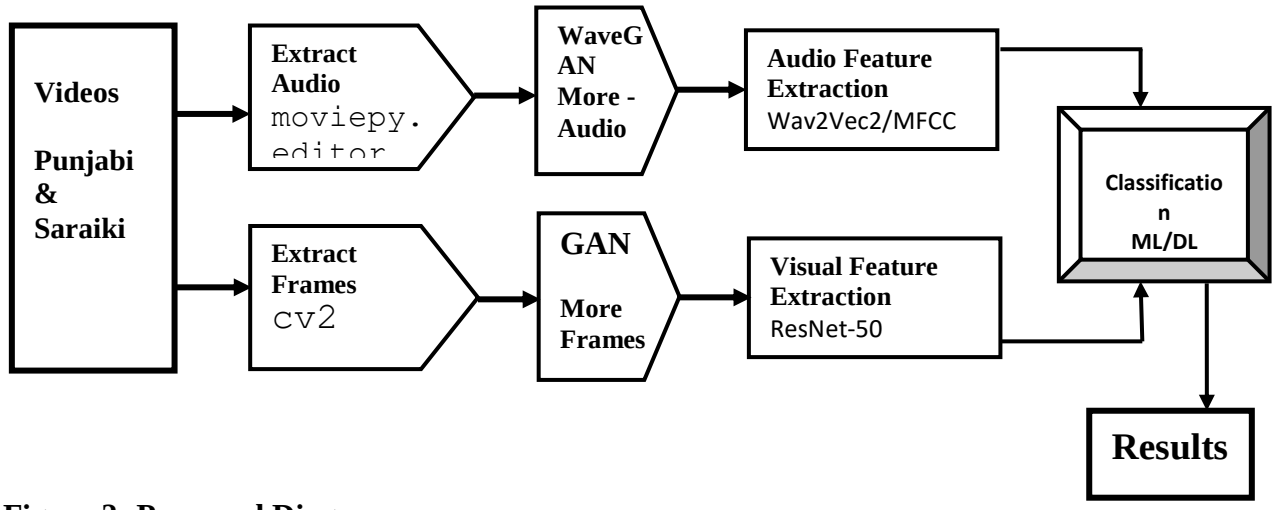

Figure 2: Proposed Diagram

4. Experiments and Results

After defining the methods, we can discuss experiments that can be used on datasets. We can use the different tools and platforms for the experiments on the dataset. In this chapter, we will perform the experiments on the dataset with all simulations and provide complete details of the experiments and results. This study is aimed at testing the effectiveness of the proposed multimodal sentiment analysis framework founded on generative artificial intelligence (AI) on the Punjab and Saraiki video data sets. The Google Colab platform, which offered an effective setting for deep learning simulations and large-scale data processing, is used to implement all experiments using Python. As part of the preprocessing process of the dataset, the audio tracks were extracted from all videos of both the Saraiki and Punjabi datasets using the MoviePy package to capture speech-based sentiment indicators. The OpenCV (cv2) package was also applied to get 20 representative frames per video to maintain visual expressions and contextual information to do further analysis. This research enhanced the data set and solved the problem of limited resources for the application of Generative Adversarial Networks (GANs) to generate additional artificial sounds and frames following preprocessing.

4.1. Experimental Data:

The experimental data in the present research are two manually developed video datasets in two languages (Saraiki and Punjabi), which were labeled into 3 sentiment categories, namely neutral, negative, and positive. The Punjabi dataset consisted of 310 videos (87 positive, 77 negative, and 146 neutral), and the Saraiki dataset consisted of 155 movies (60 positive, 43 negative, and 52 neutral). To extract both audio and video sentiment cues, MoviePy was applied to extract the audio track of all videos, and OpenCV (cv2) was applied to extract 20

frames of each video. To address the issue of data insufficiency in low-resource languages, Generative Adversarial Networks (GANs) were used to generate more synthetic data in both audio and visual forms. This improved dataset enhanced the quality of representation of sentiment classes and increased the duration of the experiment framework. The experimental data provide a good basis to make simulations to be conducted to analyze the performance of machine learning and deep learning models in multimodal sentiment analysis based on combining real and artificial samples across modalities.

4.2. Training, Testing, and Validation:

The datasets were separated into testing, validation, and training sets to guarantee unbiased model assessment. In order to preserve class balance, deep learning models used a70% training, 15% validation, and 15% testing split, whereas machine learning models used k-fold cross-validation. Wav2Vec2 was optimized on audio signals and ResNet-50 on video frames; GAN-generated data was added to increase robustness. Stable convergence was attained, and overfitting was avoided by employing strategies including early halting, dropout, and Adam optimizer with learning rate scheduling. The test set was used to measure final performance using F1-score, AUC-ROC, accuracy, precision, and recall.

4.3. Experiments:

We have performed our experiments in Python in Google Colab. All the experiments have been performed on the dataset.

4.4. Experimental flow:

To assess multimodal sentiment analysis on the Saraiki and Punjabi datasets, the experimental flow of this study was methodically set up. Using the Punjabi dataset, the tests were first carried out in the visual modality, extracting 20 representative frames per movie and processing them through ResNet-50 for feature extraction. Both machine learning and deep learning models were then used for classification. The Punjabi dataset's audio modality was then analyzed, with features derived from Wav2Vec2 and audio tracks retrieved using MoviePy. The same classifiers were then used. Following the conclusion of the Punjabi studies, the same process was carried out again for the Saraiki dataset, beginning with frame-based tests and progressing to audio-based trials. This structured experimental flow Punjabi frames → Punjabi audio → Saraiki frames → Saraiki audio ensured that each modality and dataset was evaluated separately before combining insights for comparative analysis.

In all experiments, the Punjabi frame-based results revealed that CNN was the best-performing model with the highest F1-score, recall, and AUC, whereas SVM was the most powerful among the traditional models and recurrent architectures, like LSTM and Bi-LSTM. Bilinear

LSTMs outperformed SVM in terms of the highest F1-score and AUC, suggesting that it was the best model to use when there were sequential audio features, but SVM was the best model in terms of the highest accuracy in class divisions. When it comes to Saraiki frames, all models have shown moderate performance with SVM getting the best F1-score out of machine learning techniques, with LSTM-based models slightly beating CNN, but the overall values of AUC have been low. Lastly, the overall model performance in terms of Saraiki audio classification was poor regardless of the model; the decision tree had the highest F1-score in machine learning models, and Bi-LSTM had a slight advantage over other deep learning models, but still had low improvement. All in all, the results suggested that Punjabi data showed better performance of the model than the Saraiki data, and deep learning models showed better performance using audio inputs, especially in Punjabi.

4.4.1 Results for Punjabi Frames

For the Punjabi video dataset, a total of approximately 6,200 frames were extracted across the three sentiment classes: positive, negative, and neutral. The following are the results for ML and DL classifiers.

Table 3: Results for Punjabi Frames

Model		**F1**	**Recall**	**AUC**
Machine Learning	**SVM**	0.441106	0.440299	0.979314
	DT	0.436017	0.438166	0.575963
	Random forest	0.440166	0.440299	0.701716
Deep Learning	**CNN**	0.6637	0.6663	0.8379
	LSTM	0.3256	0.3500	0.5214
	Bi-LSTM	0.3257	0.3267	0.500

This table shows that CNN performed the best among all models for Punjabi frame sentiment analysis, achieving the highest F1-score, recall, and AUC. SVM was the strongest traditional model, while LSTM and Bi-LSTM showed weak performance on image-based inputs.

Table 4: Results for Punjabi Audio

Model		**F1**	**Recall**	**AUC**
Machine Learning	**SVM**	0.447640	0.450617	0.512389
	DT	0.411441	0.407407	0.560320
	Random forest	0.412009	0.407407	0.572732
Deep Learning	**CNN**	0.225341	0.396694	0.569451
	LSTM	0.426446	0.429752	0.641250
	Bi-LSTM	0.472818	0.471074	0.672491

In the case of Punjabi audio sentiment analysis, the Bi-LSTM setting had the best performance in terms of the highest F1-score and AUC. The best machine learning model was SVM with a

lower AUC. CNN was weak on audio data, indicating that it had low appropriateness in this modality.

Table 5: Results for Saraiki Frames

Model		F1	Recall	AUC
Machine Learning	**SVM**	0.364489	0.365	0.486804
	DT	0.324499	0.325	0.493750
	Random forest	0.319904	0.320	0.485194
Deep Learning	**CNN**	0.2555	0.3483	0.5400
	LSTM	0.3256	0.3500	0.5214
	Bi-LSTM	0.3257	0.3267	0.3267

All the models worked moderately in the Saraiki frame-based results. The F1-score of SVM was most effective compared to machine learning, and LSTM and Bi-LSTM were slightly better than CNN as deep learning models, but the values of AUC are low.

Table 6: Results for Saraiki Audio

Model		F1	Recall	AUC
Machine Learning	**SVM**	0.217317	0.330579	0.555747
	DT	0.322717	0.322314	0.491685
	Random forest	0.304549	0.31405	0.517894
Deep Learning	**CNN**	0.182972	0.351648	0.500267
	LSTM	0.182972	0.351648	0.472455
	Bi-LSTM	0.212293	0.340659	0.501610

The audio results in Saraiki are generally poor in all the models. The highest F1-score of any of the machine learning models was obtained with the Decision Tree, whereas Bi-LSTM, CNN, and LSTM in deep learning models showed improvement with little difference.

5. Results for Punjabi Frame

5.1 CNN – ROC

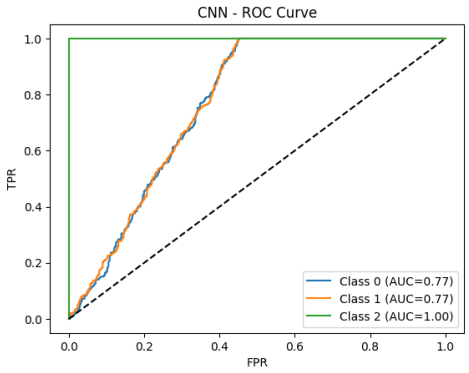

Figure 3: Shows the Results of CNN – ROC Curve, ROC curve of CNN shows results AUC of 0.838, a high ability to distinguish between classes, making it the most effective deep learning model for frames.

5.2 BILSTM ROC Curve

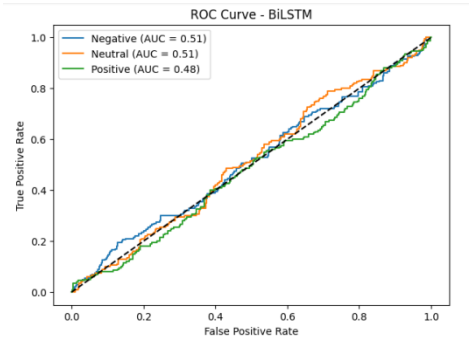

Figure 4: Shows the BILSTM ROC Curve

AUC of 0.500 shows BI-LSTM performed better than LSTM

6. Results for Punjabi Audio – SVM

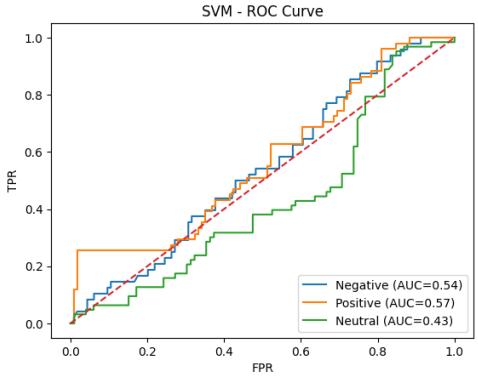

Figure 5: Shows the Results of the SVM ROC Curve, showing limited separation, with an **AUC of 0.512**.

6.1 Decision Tree –ROC Curve

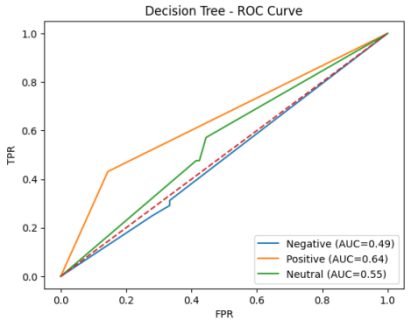

Figure 6: Shows the Results of the Decision Tree ROC Curve. Its ROC curve gave an **AUC of 0.560**, showing better separation than SVM.

6.2 Random Forest

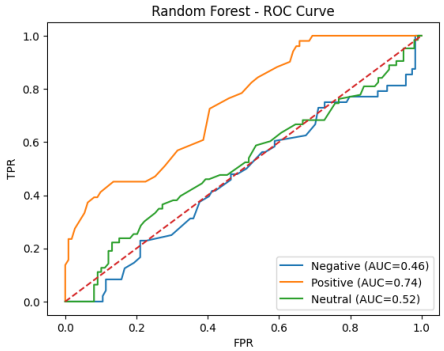

Figure 7: Shows the Results of Random Forest ROC Curve, Random Forest showed an **AUC of 0.573**, which is better than SVM and Decision Tree.

7. Punjabi audio Results for Deep Learning Classifications CNN

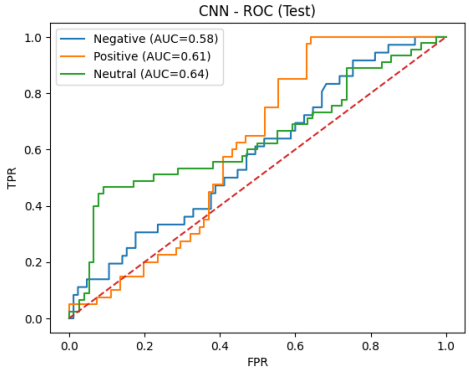

Figure 8: Shows the Results of the CNN ROC Curve

With an **AUC of 0.569**, CNN demonstrates limited ability to separate sentiment classes.

7.1 LSTM

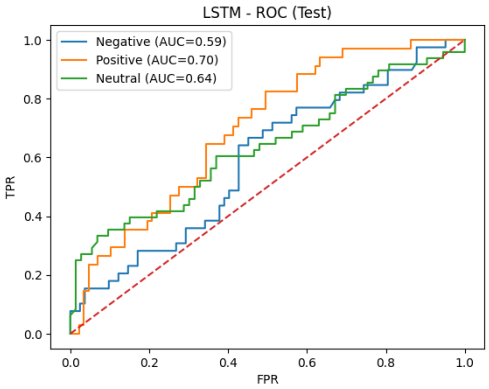

Figure 9: Shows the Results of the LSTM ROC Curve. The ROC curve for LSTM shows a fair separation ability, with an AUC of 0.641.

7.2 BILSTM ROC

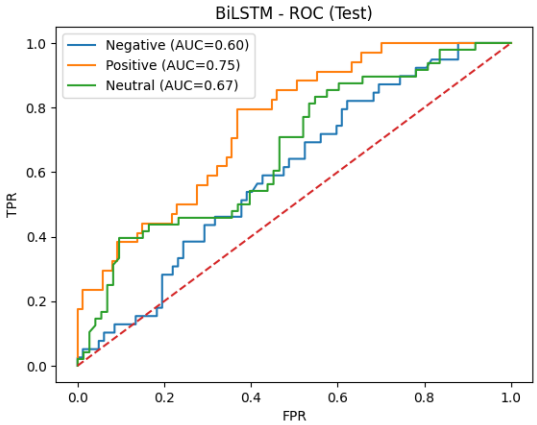

Figure 10: Shows the Results of the BILSTM ROC Curve, with an AUC of 0.672. BiLSTM shows the highest discriminative ability among the tested models.

8. Results for Saraiki Frames

8.1 CNN ROC

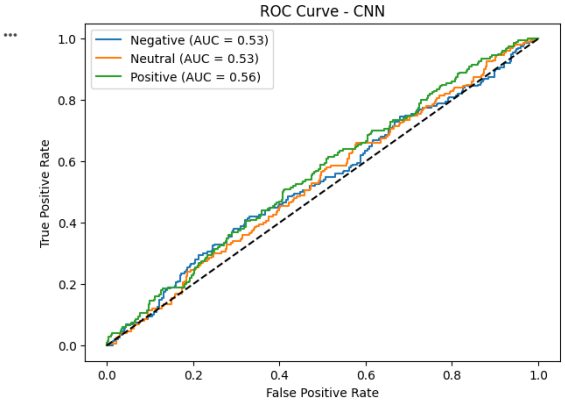

Figure 11: Shows the Results of CNN ROC Curve, the ROC curve of CNN shows an AUC of 0.5400, which is better than other classifiers.

8.2 LSTM ROC

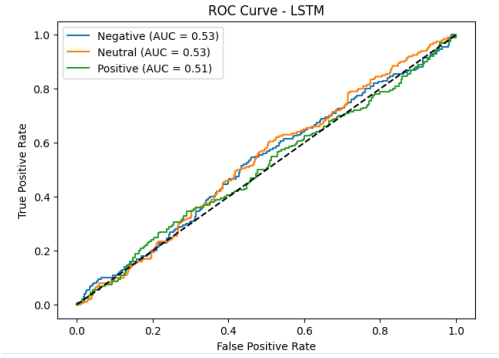

Figure 12: Shows the Results of LSTM ROC Curve, ROC curve of LSTM shows results AUC of 0.521, which is lower than other CNNs.

9. Results for Saraiki Audio – SVM

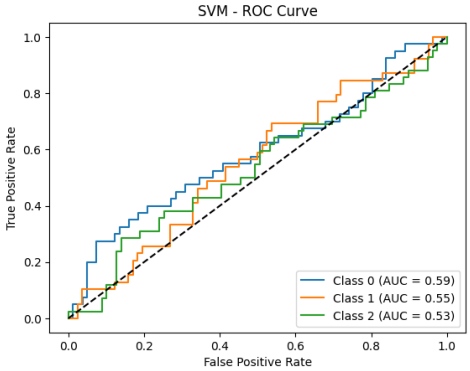

Figure 13: Shows the Results of SVM ROC Curve, the ROC curve of SVM shows an AUC of 0.555, which is better than other DT and Random Forest.

9.1 Random Forest ROC

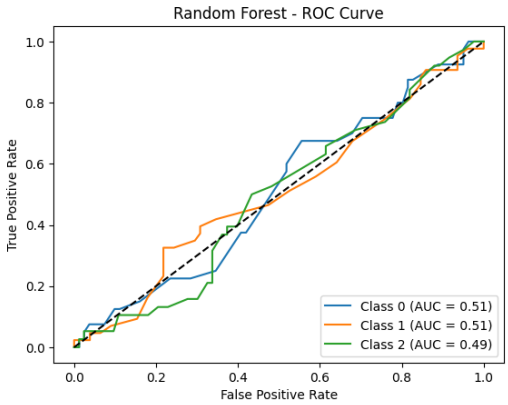

Figure 14: Shows the Results of Random Forest ROC Curve, the ROC curve of Random Forest shows an AUC of 0.5178, which is better than that of the Decision Tree.

9.3 CNN ROC

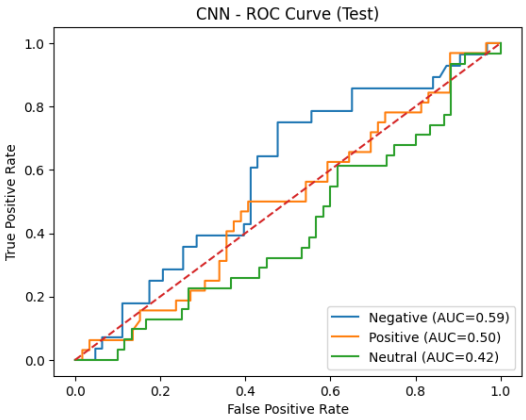

Figure 15: Shows the Results of CNN ROC Curve, the ROC curve of CNN shows an AUC of 0.500, which is better than other classifiers.

9.4 BILSTM ROC

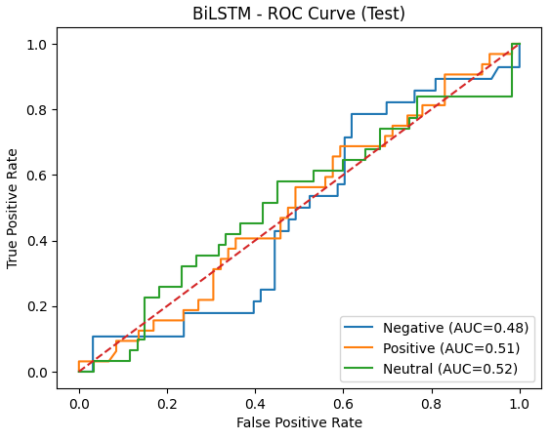

Figure 16: Shows the Results of BILSTM ROC Curve, ROC curve of BI-LSTM shows results AUC of 0.501, which is higher than LSTM classifiers.

10. Discussion

The experiment findings clearly showed that DL classifiers outperformed traditional ML models in both Punjabi and Saraiki datasets. The deep learning models that performed the most promising were CNN and BiLSTM, with the F1-scores of 0.66 and 0.47, respectively, and the AUC values of 0.83 and 0.67, respectively. These findings suggest that the models were very useful in describing the spatial and temporal sentiment information of the visual frame and audio cues. Conversely, other ML classifiers, including SVM, Random Forest, and Decision Tree, had relatively lower F1-scores, which points to their inability to process multimodal data that can be complex. Data augmentation using GAN was quite important in balancing out the classes, as

well as the generalization of the model, whereas the ResNet-50 and Wav2Vec2 feature extraction were again useful in improving performance. Generally, integrating generative augmentation with deep learning feature extraction statistically significantly improved the F1-score and the AUC, which validates the effectiveness and viability of the suggested framework in low-resource language sentiment analysis.

11. Future Work/Conclusion

Chapter 1 introduced the fundamental issue of sentiment analysis in low-resource languages like Punjabi and Saraiki, emphasizing the importance of addressing the lack of annotated datasets and the difficulties faced in multimodal analysis. Here, we offer a comprehensive discussion of the research's findings and overall flow, referencing the earlier chapters for consistency.

The overall conversation of this Study begins with Chapter 1, which introduces the problem area and the significance of the study. Sentiment analysis is challenging for Punjabi and Saraiki because they lack significant annotated datasets and scholarly attention, in contrast to high-resource languages like English. Additionally, it emphasized the value of multimodal analysis by combining video and audio elements to produce a more realistic sentiment representation. Using generative AI and sophisticated DL models to close this gap was one of the well-defined goals. This chapter offers a thorough analysis of the research's conclusions and overall flow, making connections to the previous chapters to maintain coherence.

The literature review in Chapter 2 was also comprehensive, and it examined the volume of research on audio and video-based sentiment analysis. The gaps in the research included the necessity of a better classification accuracy, the lack of the use of generative methods, and the unavailability of resources. This formed the basis of the methodology, which was explained in Chapter 3 and involved the use of several techniques to address these limitations. With the assistance of Generative Adversarial Networks (GANs) in creating additional frames and sample audio, the lack of data was addressed. The extraction of rich features in raw audio data was carried out using Wave2Vec, and linguistic and auditory characteristics were maintained. ResNet-50 was applied to video frames in order to extract features, and this resulted in the development of a deep visual representation. Moreover, to compare the performance of sentiment classification among traditional and state-of-the-art approaches, the performance was measured with Deep Learning (DL) models and with the results of the Machine Learning (ML) classifiers.

The tests and findings were, in turn, reported in Chapter 4, where systematic analyses of data (Saraiki and Punjabi) revealed the extent to which the proposed multimodal framework was

successful. Chapter 4, in the light of the tests and discoveries, showed the efficiency of the suggested multimodal architecture in terms of detailed testing on the Saraiki and Punjabi datasets. The results established that the GAN-based augmentation enhanced the diversity of datasets, ResNet-50 enhanced framework-level prediction of sentiment, and Wave2Vec offered stable audio sentiment classification. Moreover, DL models never stop to achieve better F1-score, recall, accuracy, and precision compared to ML methods. Altogether, the discussion demonstrates that the generative AI techniques and multimodal deep learning models were effectively applied to fill the research gaps revealed in the first chapters and significantly improved sentiment analysis in under-resourced languages. The experiment findings clearly show that DL classifiers performed better than traditional ML models in both Punjabi and Saraiki datasets. The deep learning models that performed the most promising were CNN and BiLSTM, with the F1-scores of 0.66 and 0.47, and AUC values of 0.83 and 0.67, respectively. This research had a detailed architecture of multimodal sentiment using AI generation. study and the language of low resources, including Punjabi and Saraiki. The study used Wave2Vec to directly extract linguistic and audio patterns from raw audio signals, ResNet-50 to extract discriminative features from video frames, and Generative Adversarial Networks (GANs) to overcome data scarcity by producing more audio samples and video frames. For performance evaluation, machine Learning and Deep Learning classifiers were used, where deep learning, over and over again, outperformed traditional ML methods. The experimental findings verified that classification recall accuracy, precision, and F1-score were all markedly improved by combining generative models with multimodal feature extraction. These results not only show how successful the suggested method is, but they also offer a framework for future research into sentiment analysis in underrepresented languages. There are a number of directions for further research despite these encouraging findings. First, increasing the quantity and diversity of the dataset to include speakers of Punjabi and Saraiki across a wider range of dialects, age groups, and speech patterns may improve the generalization and resilience of the model. Furthermore, cross-lingual transfer learning could be explored to extend the framework to other low-resource languages beyond Punjabi and Saraiki, broadening its societal and academic impact. In conclusion, this research successfully addressed the identified gaps by combining generative AI with multimodal sentiment analysis, offering a novel framework that advances the study of low-resource languages. The results provide strong evidence that generative augmentation, deep feature extraction, and multimodal fusion can substantially improve sentiment classification accuracy. However, the work also opens up exciting opportunities for future exploration,

ensuring that this study serves as both a meaningful contribution and a foundation for continued research in the field of multimodal sentiment analysis.

To improve the results of Saraiki, there is a need to generate more data is also be included in our future work.

Reference

[1]. Abdu, S. A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion , 204-226.

[2]. Ahmed, N. A. (2024). A Novel Approach for Sentiment Analysis of a Low Resource Language Using Deep Learning Models.

[3]. Alderazi, F. A. (2024). Generative artificial intelligence in topic-sentiment classification for Arabic text: a comparative study with possible future directions. PeerJ Computer Sci .

[4]. Aslam, A. S. (2023). Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. Applied Soft Computing .

[5]. Bagchi, A. (2025). Introduction to pandas. In Essentials of python for artificial intelligence and machine learning. Cham: Springer Nature Switzerland. , pp. 161-196.

[6]. Bashir, M. F. (2023). Context-aware emotion detection from low-resource urdu language using deep neural network. ACM Transactions on Asian and Low-Resource Language Information Proc .

[7]. Bhavani, K. (2025). AI Verse-A Multimodal Generative Platform. Journal of Nonlinear Analysis and Optimization .

[8]. Brennan. (2024, NoV 26). random-forest-algorithm. Retrieved SEP 28, 2025, from builtin.com: https://builtin.com/data-science/random-forest-algorithm

[9]. Cesar. (2024). A Systematic Review on Responsible Multimodal Sentiment Analysis in Marketing Applications. IEEE .

[10]. Das, R. &. (2023). Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. ACM Transactions on Asian and Low-Resource Language Information Processing. 1-30.

[11]. Faria, D. R. (2024). Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning. Applied Sciences .

[12]. Firdaus, M. C. (2020). EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. IEEE Transactions on Affective Computing. IEEE , 1555-1566.

[13]. Gandhi, A. A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion .

[14]. gitee. (2021). ParallelWaveGAN. Retrieved 10 28, 2025, from https://gitee.com/: https://gitee.com/copperdong/ParallelWaveGAN#https://gitee.com/link?target=https%3A%2F%2Farxiv.org%2Fpdf%2F1910.11480.pdf

[15]. Gokani, M. (2023). Exploring Sentiment Analysis in Low-resource Languages . (Doctoral dissertation, International Institute of Information Technology Hyderabad).

[16]. Goodfellow, I. (2020). Generative adversarial networks. Communications of the ACM , 139-144.

[17]. Han, W. C. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 international conference on multimodal interaction (pp.

[18]. Harris, C. (2020). Array programming with NumPy. 357-362.

[19]. Holdsworth, J. (2024, June 17). Retrieved December 4, 2024, from www.ibm.com: https://www.ibm.com/topics/deep-learning

[20]. Hossain, E. S. (2022). MemoSen: A multimodal dataset for sentiment analysis of memes. (pp. 1542-1554).
[21]. Hwang, J. (2023). Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch. IEEE .
[22]. Ibm. (2023, december 23). upport-vector-machine. Retrieved 10 28, 2025,
[23]. Imambi, S. (2021). In Programming with TensorFlow: solution for edge computing applications . Cham: Springer International Publishing. , PyTorch.87-104.
[24]. Jabeen, S. L. (2023). A review on methods and applications in multimodal deep learning. ACM Transactions on Multimedia Computing,Communications and Applications .
[25]. Juba, B. (2019). . Precision-recall versus accuracy and the role of large datasets. pp. 4039-4048.
[26]. Kamal, K. (2024). An Evaluation of Multi-Modal Approaches for Sentiment Analysis using Deep Learning. Doctoral dissertation, Auckland University of Technology .
[27]. Khan, L. A. (2021). Urdu sentiment analysis with deep learning methods. IEEE access .
[28]. Khan, R. A. (2024). Sentiment Analysis Using Machine Learning with Feature Extraction Methods on Resource-Deprived Language.
[29]. Koshikawa. (2025). Evaluation of Different Training Strategies and Recognizers in Low Resource Speech Recognition Using Wav2vec2. In International Conference on Machine Learning and Computing , 508-518.
[30]. Krugmann, J. O. (2024). Sentiment Analysis in the Age of Generative AI. Customer Needs and Solutions .
[31]. Kumar, V. (2022). Multimodal sentiment analysis using speech signals with machine learning techniques. IEEE .
[32]. Li, B. (2021). Facial expression recognition via ResNet-50. International Journal of Cognitive Computing in Engineering , 57-64.
[33]. Liu, J. L. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean Journal of Basic and Applied Sciences (MJBAS) , 243-260.
[34]. Malviya, S. T. (2020). Machine learning techniques for sentiment analysis: A review. A Journal of Physical Sciences, Engineering and Technology .
[35]. matias, M. (2022). Machine learning approach for topic and sentiment analysis in multilingual opinion and low source languages. School of technology .
[36]. Mercha, E. M. (2023). Machine learning and deep learning for sentiment analysis across languages. Neurocomputing .
[37]. Muthusamy, S. (2015). Computer generated summaries keyword extraction from video content using NLP techniques.
[38]. Naidu, G. (2023). A review of evaluation metrics in machine learning algorithms. In Computer science on-line conference. Cham: Springer International Publishing. , pp. 15-25.
[39]. Parveen, S. (2021). IEEE. A motion detection system in python and opencv , 1378-1382.
[40]. presidio. (2022, 06 12). The-power-of-generative-adversarial-networks. Retrieved 09 23, 2025, from www.presidio.com: https://www.presidio.com/technical-blog/exploring-the-power-of-generative-adversarial-networks-gans-with-azure/
[41]. Qaiser, S. (2021). A comparison of machine learning techniques for sentiment analysis. . Turkish Journal of Computer and Mathematics Education , 1738-1744.
[42]. Qamar. (2019). IEEE. 66-70.
[43]. Reddy, P. (2023). Enhancing Audio Synthesis with WAVEGAN: A Generative Adversarial Network (GAN) Approach. In International Conference on Information and Management Engineering , 107-113.

[44]. researchgate. (2025). Workflow-of-ResNet-50-architecture_fig4_367504190. Retrieved 10 28, 2025, from www.researchgate.net: https://www.researchgate.net/figure/Workflow-of-ResNet-50-architecture_fig4_367504190

[45]. Ruijie, H. (2024). Leveraging Librosa for Speech Emotion Recognition: Techniques and Applications. IEEE , pp. 245-249.

[46]. Sathyanarayanan. (2024). Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research , 4023-4031.

[47]. Scapicchio, M. (2024, March 22). Retrieved December 4, 2024, from www.ibm.com: https://www.ibm.com/topics/generative-ai

[48]. sciencedirect. (n.d.). bidirectional-long-short-term-memory-network. Retrieved 10 28, 2025, from www.sciencedirect.com: https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network

[49]. Sharma. (2018). Multimodal sentiment analysis using deep learning. IEEE , 1475-1478.

[50]. Sharma, R. F. (2018). Multimodal sentiment analysis using deep learning. IEEE , 1475-1478.

[51]. Sial, A. (2021). Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python. International Journal , 277-281.

[52]. Stryker, C. (2024, august 11). Retrieved December 4, 2024, from www.ibm.com: https://www.ibm.com/topics/natural-language-processing

[53]. Stryker, C. (2024, july 15). Retrieved December 4, 2024, from ibm: https://www.ibm.com/think/topics/multimodal-ai

[54]. Sun, L. X. (2021). Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model . 15-20.

[55]. Tafvizi, A. (2022). Attributing auc-roc to analyze binary classifier performance.

[56]. Thakkar, G. H. (2024). M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets.

[57]. Uppari, R. (2020). Comparison between KERAS library and FAST. AI library using convolution neural network (image classification). Dublin Business School) .

[58]. Weber, M. (2021). A tensorflow library for deep labeling. arXiv preprint arXiv. Deeplab2 .

[59]. wikipedia. (n.d.). Generative AI.

[60]. Xu, N. &. (2017). Multisentinet: A deep semantic network for multimodal sentiment analysis. 2399-2402.

[61]. Yadav, A. &. (2023). A deep multi-level attentive network for multimodal sentiment analysis. ACM Transactions on Multimedia Computing , 1-19.

[62]. Yu, W. X. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence.

[63]. Zoumana. (2013, Nov 14). introduction-to-convolutional-neural-networks-cnns. Retrieved Sep 28, 2025, from www.datacamp.com: https://www.datacamp.com/tutorial/introduction-to-convolutional-neural-networks-cnns