Emotion Detection From Urdu Speech

Shayaan Qazi
Computer Science
Habib University

Ikhlaas Ahmed
Computer Sceince
Habib University

Sameer Kamani
Computer Sceince
Habib University

Hunain Abbas
Computer Sceince
Habib University

***Abstract*—With the increased use of AI, the field of sentiment analysis has also grown. Identifying emotions accurately from speech has many applications today, ranging from mental health to customer service. Urdu presents unique challenges in speech-based emotion recognition due to its complex nature and lack of available resources. In this paper, we plan on a deep learning-based approach to detect emotions from Urdu speech. By using the latest models and an extensive dataset, we plan to develop a model that performs well and contributes to the world of sentiment analysis, especially for Urdu.**

***Index Terms*—emotion detection, sentiment analysis, Urdu speech, deep learning, audio classification, natural language processing.**

I. Introduction

Human emotions are an important part of communication, conveyed not just through words but also through voice. Use of AI to identify such emotions is an emerging field, especially in natural language processing (NLP), audio analysis and AI. Audio sentiment detection is the task of analysing and classifying emotions by monitoring how speech sounds interact with emotion.

Urdu is our national tongue and holds a lot of importance in our daily lives. It has a rich and diverse culture, and boasts some impressive poetry. However, due to Urdu having a complex nature and there being lack of available resources, especially high-quality audio, there has not been significant work done on sentiment analysis in Urdu. Also, its difficult vocabulary and diverse dialects only make our work even harder.

That being said, deep learning has now become a powerful tool for sentiment analysis. Such models are not only a lot more accurate but also work well on many different languages, Urdu being one of them. This project aims to use that to develop an accurate model to detect sentiments in Urdu speech and contribute something meaningful to our native language.

II. Research Question

The research question we are addressing is: "How can we accurately detect emotions and sentiments in Urdu speech using deep learning?" Specifically, we are trying to solve the problem: "Given an Urdu audio clip, how accurately can we classify the speaker's emotion?"

This work is important because understanding human emotions has become crucial in many industries. In customer service, mental health, marketing, and public relations, knowing how people feel can help organizations make better decisions. For example, emotion detection allows AI chat bots to serve as virtual therapists, without the discomfort of seeing a therapist in person or confiding in a family member.

Furthermore, it is useful for businesses to quickly and accurately understand customer feedback. Unlike current methods such as surveys and polls, automated sentiment detection allows businesses to monitor real-time feedback from sources like customer reviews and support calls. This not only speeds up the process but also gives more useful information.

Although sentiment analysis has been worked on extensively in many languages, there is very little research on Urdu. Our work aims to address this gap by using deep learning techniques. By developing accurate and reliable models for detecting emotions in Urdu speech, we also hope to contribute to the broader field of sentiment analysis and provide valuable tools for researchers and professionals, particularly who are working on Urdu.

III. Literature Review

A study by Siddique Latif [1] introduced the first spontaneous Urdu speech emotion database which had 400 sounds from talk shows that represented emotions such as anger, happiness, sadness and neutral. This database was compared with datasets of Western languages which included EMO-DB which was German, SAVEE using English and EMOVO which had Italian. The Urdu database contained acted emotions. The classification of emotions into positive or negative categories was performed using Support Vector Machines (SVM). The speech features such as energy and frequency were extracted with the "openSMILE" toolkit. The baseline results showed that SVM did better than other classifiers like Logistic Regression and Random Forest. However a major drop in accuracy was seen in cross language testing due to data differences. At least 30% of the test data in the training phase improved accuracy across languages. Multi language training involved a mix of Western language datasets and enhanced performance but still fell short of the baseline accuracy achieved with individual datasets. The study showed that while cross lingual emotion recognition is feasible for Urdu more advanced machine learning or deep learning models are necessary. The study's future work will focus on exploring deep learning techniques for feature extraction to further enhance cross lingual emotion recognition.

In another study, Badriyya B. Al-onazi [2] proposed a transformer based approach for multi language speech emotion recognition mainly focusing emotions in Arabic which has been unexplored. The model was trained and tested on four

publicly available datasets: BAVED which was Arabic, EMO-DB which was in German, SAVEE which had British English and EMOVO with Italian. All datasets contained a variety of emotions. Data augmentation techniques included noise addition and time stretching. These were used to increase the size of the training dataset and improve model performance. A total of 273 features were analyzed from each audio to capture important details such as pitch and tone. The model used a transformer architecture which used self attention mechanisms to find complex patterns within data. Evaluation results showed that the model achieved 95.2% accuracy on the BAVED dataset, 93.4% on EMO-DB, 85.1% on SAVEE, and 91.7% on EMOVO. The transformer model performed better than the traditional methods like CNN and SVM and showed major improvements in emotion recognition across all datasets through effective data augmentation and advanced feature extraction techniques. The study showed that this approach could lead to advancements in recognizing emotions in multiple languages as Arabic and suggested future research could explore integrating RNNs with transformers and applying the method to languages with limited data.

Riyaz Shaik [3] and his team used Automatic Speech Recognition (ASR). Their study was useful for dealing with Urdu's complex nature as they identified that and talked about ways to address it. They used a custom made dataset, which had 600 Urdu audio clips, categorized into three emotions: positive, negative and neutral. They used noise removal techniques while preprocessing their data and removed silence from the audio clips. Important features, such as energy patterns and frequency patterns were extracted in the feature extraction stage. The study used a Hidden Markov Model also known as HMM to recognize temporal patterns in speech and Dynamic Time Warping also known as DTW to compare test speech with known emotional words while taking care of variations in speech speed. By using both HMM and DTW models the accuracy of emotion recognition was improved even in the presence of background noise. The proposed approach achieved an unbelievable accuracy of 97.1% by adding various speech features and techniques showing a safe method for sentiment analysis in Urdu speech.

Umar Farooq [4] did his work by introducing a multimodal sentiment analysis approach and by incorporating text, audio and visual data to better capture sentiment in Urdu. A custom dataset of 44 YouTube videos were compiled with 1372 segmented clips for sentiment polarity. The study used Bidirectional Long Short Term Memory also known as BLSTM for text analysis and 3D Convolutional Neural Networks also known as CNNs for audio visual data and he achieved a 95.35% accuracy rate. The combination of audio and text features proved very effective and they did better than unimodal models. This research highlights the potential of multimodal approaches in sentiment analysis though the authors acknowledged limitations in dataset size and suggested exploring transformer models in future work.

Bilal Khan [5] introduced an Urdu speech collection for emotion recognition which plays an important role in human computer interaction. The dataset had recordings from 20 native Urdu speakers who expressed 5 emotions which included sadness, happiness, neutral, disgust and anger. They used feature extraction methods which included Mel Frequency Cepstral Coefficients also known as MFCCs and Linear Prediction Coefficients also known as LPC alongside machine learning models like K Nearest Neighbors also known as kNN, Random Forest and SVM. While kNN performed best with an accuracy of 76.5%. The study highlighted the difficulty in classifying more complex emotions like "disgust." Despite limitations such as the small dataset size. The study still manages to lay down the groundwork for future research by recommending a shift to deep learning models to improve classification accuracy. He also recommended to train on other deep learning models such as CNN and RNN to improve classification accuracy.

In a study by Marium Mateen and Narmeen Zakaria Bawany [6] Deep Learning Approach for Detecting Audio Deepfakes in Urdu focused on the growing threat posed by AI-generated fake voice recordings, especially in languages like Urdu, which is majorly neglected in existing research. The study worked on 400 audio clips; the real ones were collected from various online sources while the fakes were made using the "Real Time Voice Cloning" (RTVC) tool. They developed a deep learning model using LSTM and achieved 91% accuracy on their dataset. Despite this success, they talked of a need to use larger datasets and possibly more advanced techniques, such as transformers and autoencoders.

In another study Ashraf Ullah and others [7], they conducted research on Threatening Language Detection from Urdu Data with a Deep Sequential Model which detected threatening language on social media platforms in Urdu. The authors created a dataset of 3564 manually-labeled Urdu tweets categorized as either threatening or non threatening. They then applied data augmentation techniques like back translation to increase the dataset size to 7128 tweets. The research focused on preprocessing steps which included cleaning the data, creating a stop words list and stemming dictionary for Urdu which improved the model's ability to detect threats. They developed an LSTM model which got an accuracy of 81.96%, which is better than traditional models such as SVM and MLP. We learned that getting a good accuracy requires a decent-sized dataset and refinement of preprocessing steps before training the model.

IV. Data Set

In this section we will discuss the dataset, its characteristics, and it's acquisition. It offers detailed information about the audio files used for training, testing and validating our emotion recognition models.

A. Summary

Our dataset consists of about 14,000 audio files in WAV format, divided into four distinct emotions: Anger, Happiness, Neutral, and Sadness. The dataset has been sourced from SEMOUR+, an online Urdu audio dataset.

The collection comprises of voice recordings from 24 actors, each saying from a list of 235 words in 4 different emotions. The first eight actors contributed 235 recordings for each emotional category, while actors 9 to 24 provided 100 recordings per emotion, creating a diverse dataset for emotion recognition.

B. Acquisition

Our primary source is SEMOUR+ dataset. The SEMOUR+ dataset is built on the SEMOUR dataset, which we accessed through a literature review of "An Urdu Speech Corpus for Emotion Recognition". SEMOUR+ provides voice notes that showcase each actor conveying one of four emotions. The data acquisition process required a comprehensive search to access the SEMOUR+ dataset online.

C. Demographics:

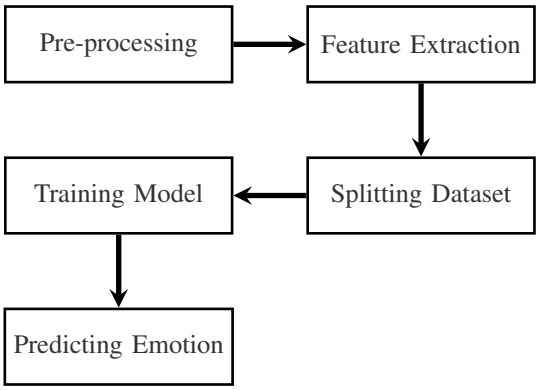
The demographics of the actors are as follows:

TABLE I
DEMOGRAPHICS OF VOICE ACTORS

No	Actor	Gender	Age Range
1	Actor 1	Male	20-25
2	Actor 2	Male	30-35
3	Actor 3	Male	20-25
4	Actor 4	Male	20-30
5	Actor 5	Female	20-25
6	Actor 6	Female	20-25
7	Actor 7	Female	30-40
8	Actor 8	Female	30-40
9	Actor 9	Female	20-25
10	Actor 10	Female	30-40
11	Actor 11	Female	30-40
12	Actor 12	Male	30-40
13	Actor 13	Male	30-40
14	Actor 14	Male	30-40
15	Actor 15	Male	30-40
16	Actor 16	Male	30-40
17	Actor 17	Male	30-40
18	Actor 18	Male	30-40
19	Actor 19	Male	30-40
20	Actor 20	Male	30-40
21	Actor 21	Male	30-40
22	Actor 22	Male	30-40
23	Actor 23	Male	30-40
24	Actor 24	Male	30-40

V. METHODOLOGY

Our methodology consists of 5 steps: pre-processing, feature extraction, splitting dataset, training model and predicting emotion.



A. Pre-processing

We stretched both the pitch and duration of the audio to enhance feature extraction, enabling us to capture subtle frequency and timing variations that might be lost in the original recording. We can enhance certain tonal quantities by manipulating the pitch, which makes it much easier to identify the characteristics of the sound. Additionally, by extending the duration, we are able to have a more detailed view of the audio waveform, giving a more accurate analysis of its underlying patterns. These modifications ultimately improved our ability to extract more meaningful insights from the audio data.

B. Feature Extraction

We used Mel-frequency cepstral coefficients (MFCCs) for feature extraction. MFCCs are extracted from the short-term power spectrum of sound to match how humans generally perceive sound. Human ears are much more responsive to sounds in lower frequencies than those in higher ones, MFCCs account for this by applying a mel filter bank to the audio signal. This approach enhances the modeling of speech patterns, making MFCCs particularly useful for tasks like emotion recognition, and music genre classification. By lowering the complexity of audio data while keeping important details about its sound features, MFCCs improve how well different machine learning methods work in analyzing audio.

C. Splitting Dataset

We split our dataset into 3 categories: Training, Validation and Test. While splitting we applied a 10:1:3 ratio, meaning for every audio file assigned for validation, 10 were assigned for Training and 3 were assigned for Testing. Training, validation, and test datasets are essential for ensuring that a model generalizes well and performs consistently on new, unseen data. The training set is utilized to train the model, while the validation set assists in fine-tuning hyperparameters and mitigating overfitting, and the Test set evaluates the model's final performance. A 10:1:3 split ratio is effective because it strikes a balance and leaves enough data to properly test how well the model performs and learn complex patterns.

D. Training Model

We conducted an extensive analysis of our dataset by employing three distinct categories of models: machine learning models, deep learning models, and transformers. Within

the machine learning framework, we utilized Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) to establish baseline performance benchmarks. For the neural network category, we implemented Convolutional Neural Networks (CNN) and the advanced RESNET50 architecture, leveraging their capacity for deep feature extraction and classification. To further enhance our analysis, we explored state-of-the-art transformer models, specifically HUBERT (hubert-large-ls960-ft) and Wav2Vec 2.0 (wav2vec2-xls-r-300m), which are designed for speech representation and understanding. Each model was systematically evaluated to understand its effectiveness in recognizing emotions from our dataset, allowing us to identify strengths, weaknesses, and opportunities for improvement across various approaches.

E. Predicting Emotion

1) Support Vector Machine (SVM): In our project we used a Support Vector Machine (SVM) model for emotion recognition from audio recordings. We extracted feature vectors such as MelFrequency Cepstral Coefficients (MFCCs) to capture the audio's temporal and spectral characteristics. The SVM aimed to construct an optimal hyperplane in a high dimensional space to separate different emotion classes by maximizing the margin between classes. This approach improved the model's generalization capabilities and efficiently handles both linear and non linear classification tasks through the use of kernel functions.

The training process involved optimizing the hyperplane weights and the model's performance was evaluated using metrics like accuracy and the F1 score. SVMs are particularly beneficial in scenarios with limited training data due to their lower risk of overfitting compared to more complex models like Convolutional Neural Networks (CNNs). By integrating feature selection techniques SVMs improve their performance by focusing on the most relevant features for emotion classification. Overall, the defined feature representations and decision boundaries make SVMs a strong choice for analyzing audio data in emotion recognition tasks.

TABLE II
SVM ASPECTS AND DETAILS

Aspect	**Details**
Kernel	Linear
C (Regularization)	0.1
Feature Type	MFCC

2) CNN Aspects and Details: In our project we also used a Convolutional Neural Network (CNN) to recognize emotions from audio recordings. We transformed the audio data into visual representations such as spectrograms and melfrequency cepstral coefficients (MFCCs) which allowed the CNN to capture both temporal and spectral features of the audio signals. The CNN architecture we used consisted of nine layers which included three convolutional layers, three pooling layers, two fully connected layers and one output layer. This structure allowed the model to automatically extract relevant features associated with different emotions such as specific frequency ranges and temporal changes.

During the training process we optimized the CNN's weights using backpropagation and the Sparse Categorical Cross entropy loss function to minimize the difference between the predicted and actual emotional labels. We trained the model for 20 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 0.001. To improve the model's performance and prevent overfitting we applied techniques such as data augmentation, dropout and batch normalization.

TABLE III
CNN MODEL CONFIGURATION ASPECTS

Aspect	**Details**
Number of Layers	9 (3 Conv + 3 Pool + 2 FC + 1 Output)
Epochs	20
Batch Size	32
Optimizer	Adam
Learning Rate	0.001
Loss Function	Sparse Categorical Crossentropy

3) Residual Network (ResNet50): the ResNet50 architecture was also used to recognize emotions from audio recordings. We transformed the audio data into visual representations such as spectrograms and melfrequency cepstral coefficients (MFCCs) which allowed the ResNet50 model to capture both temporal and spectral features of the audio signals. The ResNet50 architecture comprises of 50 layers which included convolutional layers, batch normalization and activation functions. The layers also feature skip connections that help in learning residual mappings which preserve information across layers and improving gradient flow during backpropagation.

During the training process we optimized the ResNet50 model's weights using backpropagation and the Sparse Categorical Crosse ntropy loss function to minimize the difference between the predicted and actual emotional labels. We trained the model for 30 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 0.001. To enhance the model's performance and prevent overfitting we applied techniques such as data augmentation, dropout regularization and transfer learning from pre trained models. The final output was classified into emotional states like Anger, Happiness, Neutral and Sadness using fully connected layers.

TABLE IV
RESNET ASPECTS AND DETAILS

Aspect	**Details**
Epochs	30
Batch Size	32
Optimizer	Adam
Learning Rate	0.001
Loss Function	Sparse categorical cross entropy

4) Hubert (hubert-large-ls960-ft): In this project we used the HuBERT (Hidden-Unit BERT) model for emotion recognition from audio recordings. Unlike traditional CNNs which rely on manual feature extraction such as spectrograms or

MFCCs, HuBERT uses a transformer based architecture to learn directly from raw audio files. The training process of HuBERT involves two phases: pre training and fine tuning. During pre training, HuBERT predicts masked audio segments which is similar to how BERT predicts missing words in text. This approach allows the model to capture rich contextual information without requiring labeled data. It allows the model to identify important acoustic patterns, phonetic details and prosodic features necessary for detecting emotions.

After pre training, HuBERT undergoes fine tuning on a labeled dataset for emotion recognition. The model type we used was **hubert-large-ls960-ft**. During fine tuning, HuBERT adjusts its learned representations to classify audio into emotional categories such as Anger, Happiness, Neutral and Sadness. The fine tuning process optimizes the model's weights through backpropagation using a loss function like Sparse Categorical Cross Entropy. This process reduces the gap between predicted emotional states and actual labels. The transformer architecture of HuBERT effectively captures long range dependencies and contextual nuances in speech improving its performance in emotion recognition. Additionally, techniques such as data augmentation and regularization can further improve the model's performance and generalization capabilities which makes HuBERT a powerful tool for extracting high level features from raw audio data.

TABLE V
HUBERT ASPECTS AND DETAILS

Aspect	Details
Model Type	hubert-large-ls960-ft
Epochs	10
Batch Size	16
Optimizer	AdamW
Learning Rate	3×10^{-5}, Cosine Scheduler
Loss Function	Sparse Categorical Cross-Entropy

5) K-Nearest Neighbours (KNN): K-Nearest Neighbors (KNN) algorithm is a good choice for emotion recognition from audio recordings. Unlike Convolutional Neural Networks, KNN is a nonparametric instance based learning method that classifies data points based on the attributes of their nearest neighbors in the feature space. The process began with transforming audio recordings into appropriate feature representations such as spectrograms or melfrequency cepstral coefficients (MFCCs), which capture vital temporal and spectral information. After feature extraction KNN calculated the distance between a new audio sample and all samples in the training dataset using metrics like Euclidean distance, Manhattan distance, or cosine similarity to identify the five nearest neighbors.

The emotion of the new audio sample was classified through a majority vote among its five nearest neighbors with the choice of $k = 5$ significantly influencing performance. A small k may lead to overfitting while a larger k might oversimplify classifications. KNN does not require a traditional training phase as it retains the entire training dataset which can be computationally intensive for large datasets. To improve KNN's performance in emotion detection we used techniques such as feature scaling, dimensionality reduction and careful selection of k. Additionally, implementing weighted voting where closer neighbors exert more influence, further improved classification accuracy. Overall, KNN effectively used instance similarity in the feature space to classify emotional states based on proximity to labeled examples.

TABLE VI
KNN ASPECTS AND DETAILS

Aspect	Details
Model Type	wav2vec2-xls-r-300m
Epochs	20
Batch Size	32
Optimizer	AdamW
Learning Rate	0.00003, Cosine Scheduler
Loss Function	Sparse Categorical Cross Entropy

6) Wav2Vec2.0 (wav2vec2-xls-r-300m): Wav2Vec 2.0 model for emotion recognition from audio recordings can be the best choice. Unlike traditional approaches that rely on manual feature extraction Wav2Vec 2.0 learns directly from raw audio waveforms which improves its performance across various audio analysis applications. During pre training the model uses a large amount of unlabeled audio data to learn contextualized representations through a contrastive learning approach. In the fine tuning phase the model adapts to specific tasks by training on labeled emotional data and adding a classification head.

By using a transformer architecture Wav2Vec 2.0 captures long range dependencies within audio signals which is important for recognizing emotions that may span in different speech segments. Through self attention mechanisms in multiple transformer layers the model identifies relevant parts of the audio which helps to differentiate emotional nuances. During training we used the AdamW optimizer and Sparse Categorical Cross Entropy loss function. Benefiting from data augmentation for improved generalization and performance. The model type we used was **wav2vec2-xls-r-300m** trained for 20 epochs with a batch size of 32 and a learning rate of 0.00003, using a cosine scheduler. This setup allowed Wav2Vec 2.0 to proficiently extract complex audio patterns and contextual information and streamlining the process of emotion recognition.

TABLE VII
WAV2VEC2.0 ASPECTS AND DETAILS

Aspect	Details
Number of Neighbors	5
Learning Curve Metric	Accuracy

VI. RESULTS

In this section, we evaluate the performance of various models used for emotion recognition in Urdu speech. The evaluation metrics include **Accuracy**, **Confusion Matrix**, and **Learning Curves**.

A. Support Vector Machine (SVM)

Accuracy: 62.14%
Description: The SVM model demonstrated moderate performance, correctly classifying the emotional states in 62% of the test samples.

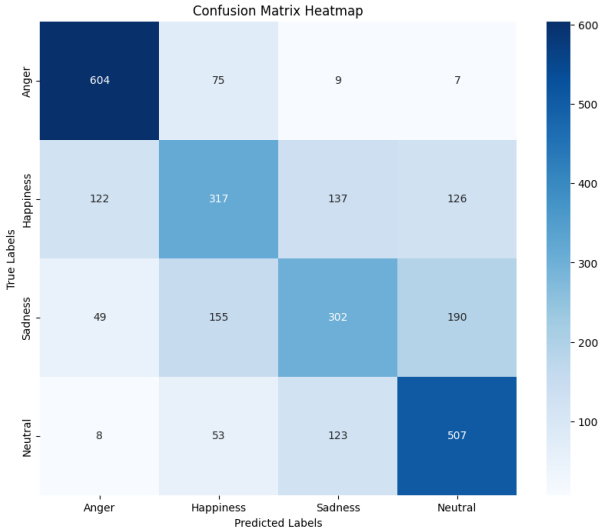


Fig. 1. SVM's Confusion Matrix Heatmap.

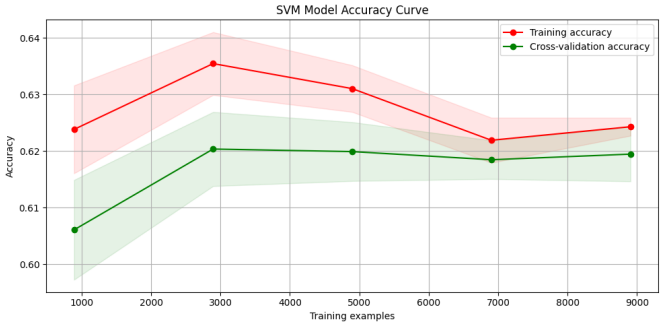


Fig. 2. SVM's Accuracy Curve.

Fig. 3. SVM's Loss Curve.

B. Convolutional Neural Network (CNN)

Accuracy: 85.09%
Description: The CNN model significantly improved the performance, correctly predicting emotions in 85% of the test samples.

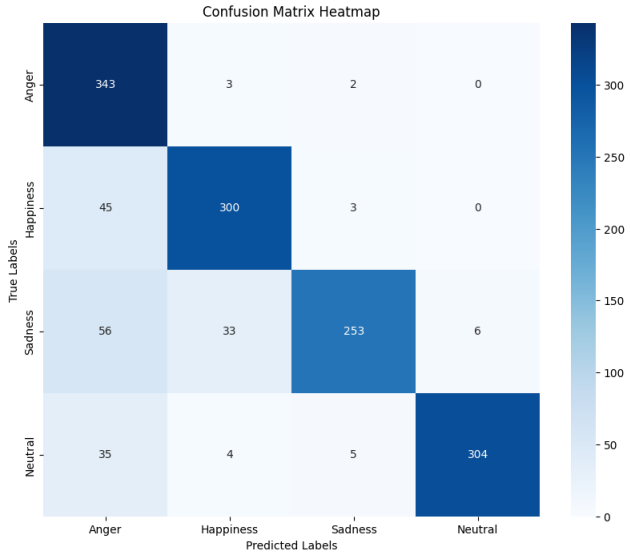


Fig. 4. CNN's Confusion Matrix Heatmap.

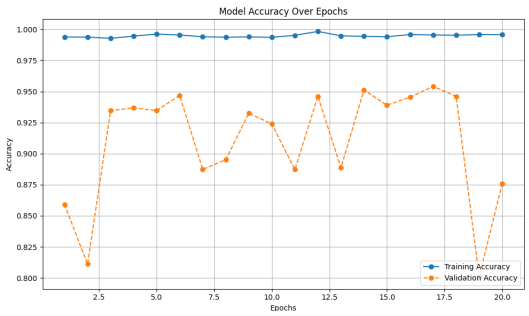


Fig. 5. CNN's Accuracy Curve.

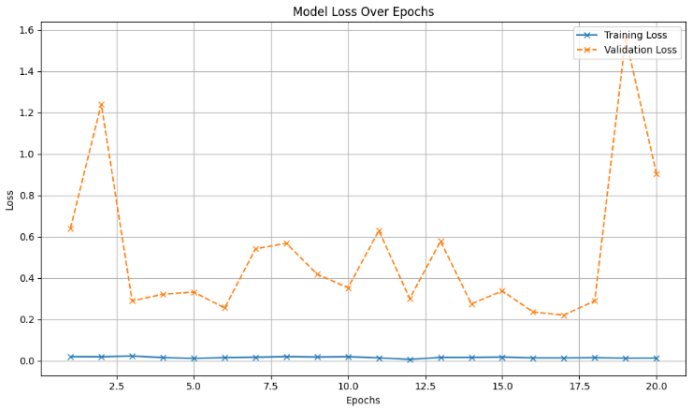


Fig. 6. CNN's Loss Curve.

C. ResNet40

Accuracy: 86.82%

Description: The ResNet40 model provided robust performance with an accuracy of 87%.

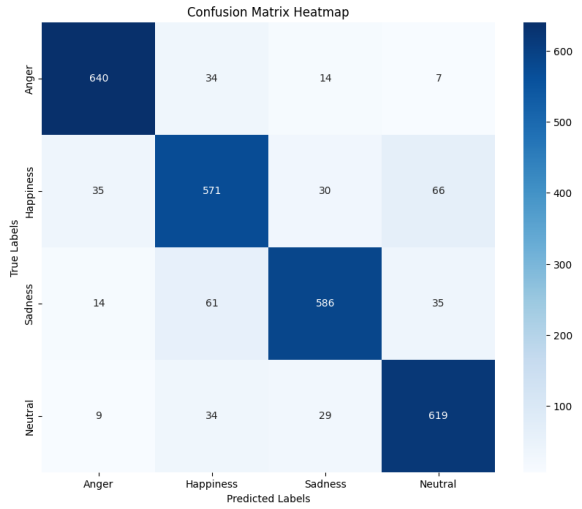


Fig. 7. ResNet40's Confusion Matrix Heatmap.

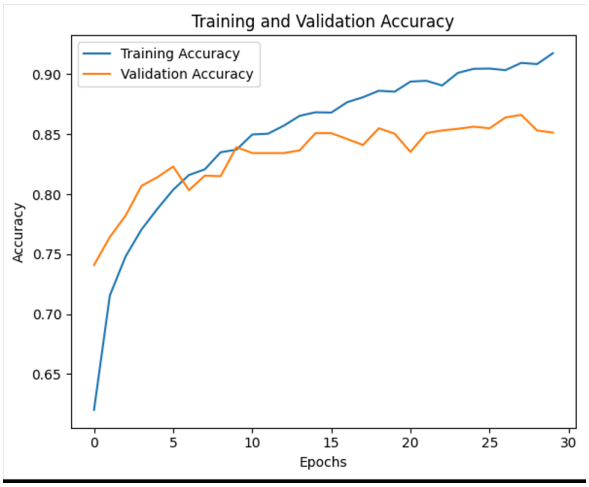


Fig. 8. ResNet40's Accuracy Curve.

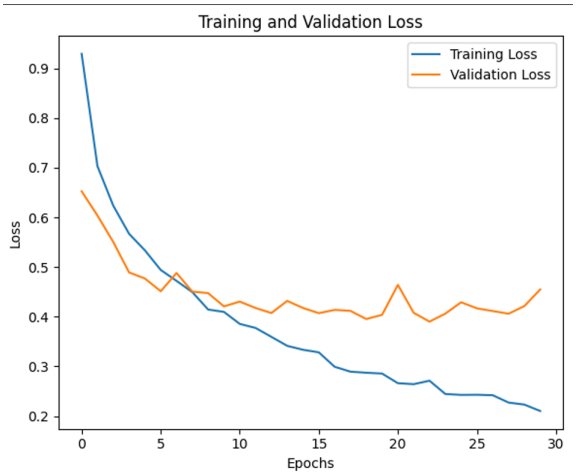


Fig. 9. ResNet40's Loss Curve.

D. HuBERT

Accuracy: 90%

Description: The HuBERT model achieved an accuracy of 90%, showing excellent performance in recognizing emotional states.

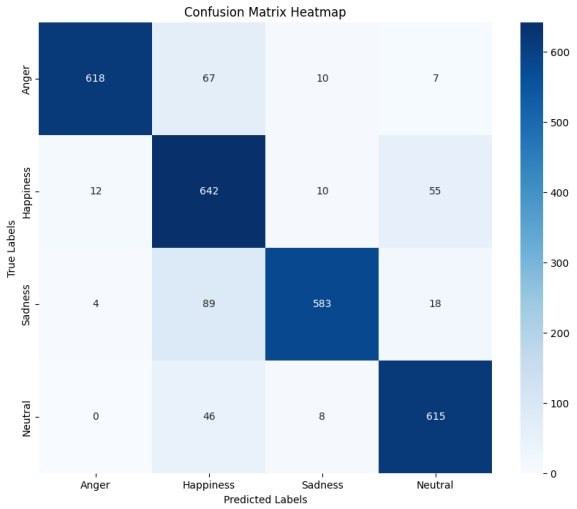


Fig. 10. HuBERT's Confusion Matrix Heatmap.

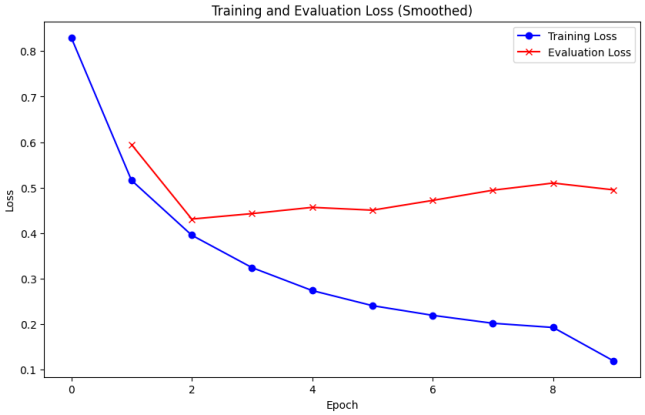


Fig. 11. HuBERT's Loss Curve.

E. K-Nearest Neighbors (KNN)

Accuracy: 91%

Description: The KNN model a good accuracy of 91%, making it one of the best models we have trained.

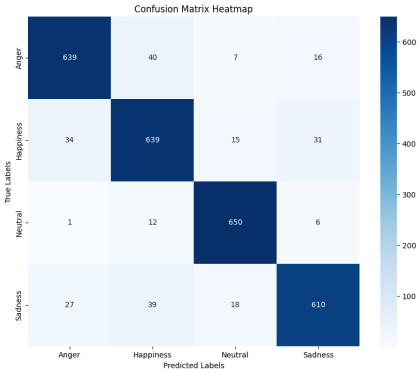


Fig. 12. KNN's Confusion Matrix Heatmap.

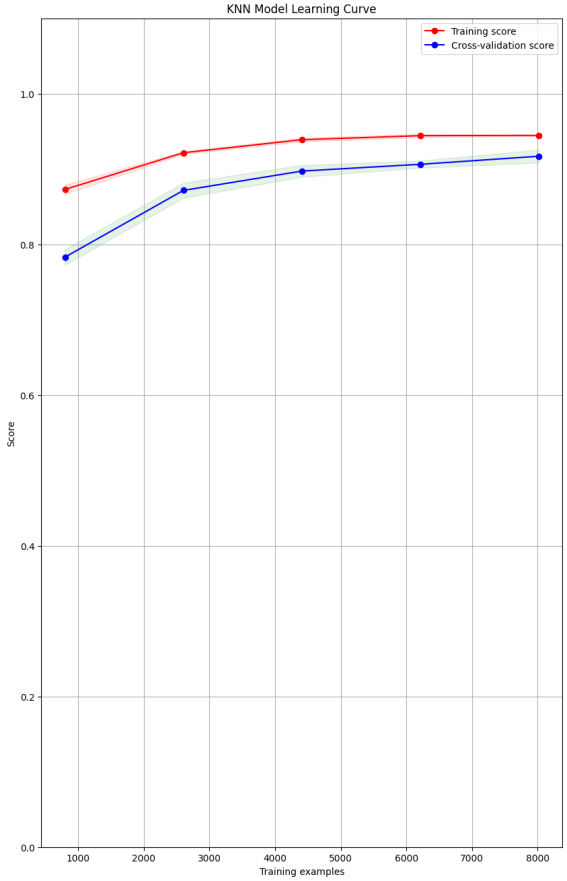


Fig. 13. KNN's Accuracy Curve.

F. Wav2Vec2.0

Accuracy: 94.50%

Description: The Wav2Vec2.0 model outperformed other models with an accuracy of 94.5%. Making this the best performing model in terms of accuracy.

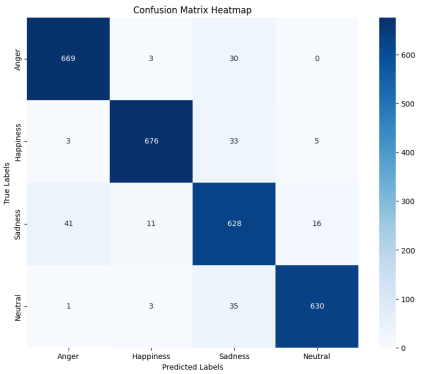


Fig. 14. Wav2Vec2.0's Confusion Matrix Heatmap.

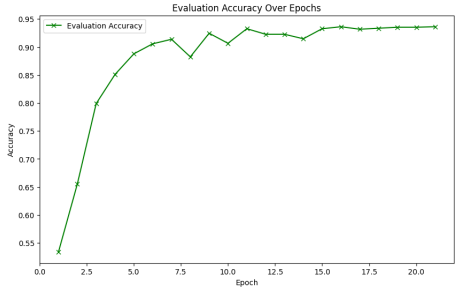


Fig. 15. Wav2Vec2.0's Accuracy Curve.

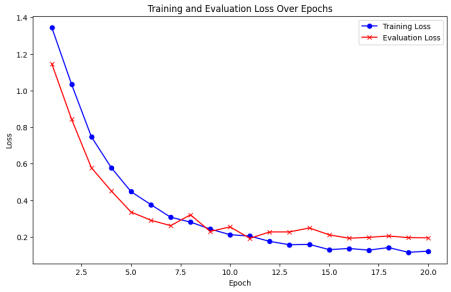


Fig. 16. Wav2Vec2.0's Loss Curve.

Our results seem as expected with the Transformers outperforming the rest of the models, followed by Deep Learning Models, and finally at the very lowest accuracy, Machine Learning Models. Although KNN did surprisingly well, outperforming HuBERT, but Wav2Vec takes the crown with an amazing accuracy of 94.50%.

TABLE VIII
COMPARISON OF MODEL PERFORMANCE

Model	**Result**
SVM	62%
CNN	85.09%
ResNet50	86.82%
HuBERT	90%
KNN	91.16%
Wav2Vec2.0	94.5%

VII. CONCLUSION

In this paper, we addressed the challenge of detecting emotions in Urdu speech by utilizing deep learning models, machine learning models, and transformers. We trained multiple models, and our findings show that among all the models, Wav2Vec2.0 has the highest the highest accuracy of 95.11%, which performed better than all other models such as KNN and HuBERT.

Wav2Vec2.0 outperforms the rest of the models due to many factors. Its extensive pre-training on unlabeled audio data allows for detailed audio recognition, while its use of transformers enhcnaces it's ability to capture relationships in audios. The model's learning approach transforms raw audio into useful information. Additionally, Wav2Vec 2.0 effectively handles variability such as background noise, and its ability to capture changes in pitch and tone, enchances emotion detection. Moreover, our large dataset allowed the model to adapt to the unique characteristics of Urdu speech, making it effective for this task.

VIII. FUTURE WORK

Future work will focus on refining our model and expanding the dataset while experimenting with various feature extraction methods to improve accuracy. We will implement different augmentation techniques and explore classifying models across a wider range of emotions and languages to enhance performance.

REFERENCES

[1] S. Latif, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8616972.

[2] [2] B. B. Al-onazi, "Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion," Applied Sciences, vol. 12, no. 18, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/18/9188.

[3] R. Shaik, "Sentiment Analysis with Word-Based Urdu Speech Recognition," Journal of Ambient Intelligence and Humanized Computing, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s12652-021-03460-x.anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] Sehar, U., Kanwal, S., Dashtipur, K., Mir, U., Abbasi, U., Khan, F. (2021). Urdu sentiment analysis via multimodal data mining based on deep learning algorithms. IEEE Access, 9, 153072-153086.

[5] Asghar, A., Sohaib, S., Iftikhar, S., Shafi, M., Fatima, K. (2022). An Urdu speech corpus for emotion recognition. PeerJ Computer Science, 8, e954.

[6] Ullah, A., Khan, K. U., Khan, A., Bakhsh, S. T., Rahman, A. U., Akbar, S., Saqia, B. (2024). Threatening language detection from Urdu data with deep sequential model. PLOS ONE, 19(6), e0290915.

[7] Mateen, M., Bawany, N. Z. (2023). Deep Learning Approach for Detecting Audio Deepfakes in Urdu. NUML International Journal of Engineering and Computing, 2(1).