Design of a System for Identification of Spoken Emotions in Urdu: Machine Learning Modeling.

Guechtouli Anis
DS Master's student
Email: guechtoulianiss7@gmail.com

Benhamdi Anis
DS Master's student
Email: benhamdianis@gmail.com

Bouhouita Hamza
DS Master's student
Email: hamza.bou2021@gmail.com

January 26, 2024

Abstract

This project aims to design a system for identifying emotions expressed in Urdu using machine learning modeling. Notably, the Neural Network model with Mel-frequency cepstral coefficients (NN MFCCs) achieved an impressive F1 score of 96.23%. The work addresses the need for culturally sensitive emotion classification tools in Urdu, advancing human-machine interactions in diverse linguistic contexts.

1 Introduction

In the dynamic realm of human-machine interactions, the intricate dance of emotions expressed by users has emerged as a key focal point. This project undertakes a nuanced exploration, aiming to design a sophisticated system crafted for the identification of three core emotions – joy, neutrality, and sadness – within spoken Urdu. Urdu, a language deeply woven into the fabric of South Asian culture, becomes the canvas upon which we paint our endeavor, addressing the scarcity of specialized emotional analysis tools tailored for this linguistic context.

Navigating the landscape of technological progress, our mission gains significance as we endeavor to create a system that authentically understands and responds to the emotional tapestry of joy, neutrality, and sadness within spoken Urdu. Our approach involves leveraging advanced machine learning models, including Neural Networks (NN), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), each trained on distinct feature representations such as Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms.

2 Materials and Methodology

The successful implementation of the proposed system for identifying spoken emotions in Urdu involves a systematic approach encompassing data collection, preprocessing, model development, training, and evaluation. The following sections detail the materials and methodologies employed throughout each phase of the project.

In the course of our research, we conducted a total of 504 experiments, each representing a unique configuration of models with distinct hyperparameter settings and variations in the size of training and testing datasets. This systematic approach allowed us to explore the impact of different factors on the performance of our machine learning models.

Data: The dataset employed in this study comprises 400 spoken utterances in Urdu. Meticulously curated, it is tailored to encapsulate the nuanced spectrum of emotional expressions inherent in spoken language. The dataset is strategically balanced, encompassing 200 instances of negative emotions, 100 instances of positive emotions, and 100 instances of neutral emotions. This intentional design ensures a comprehensive representation of emotional diversity.

To ensure a robust evaluation of the models, the dataset is divided into two main subsets: a training set, employed for model development, and a testing set, used for rigorous evaluation. This composition facilitates a comprehensive exploration of the machine learning models' ability to identify and classify various emotional states in spoken Urdu expressions.

For experimentation, the dataset is further split into the following proportions: 20% test, 80% training; 25% test, 75% training; 30% test, 70% training; 35% test, 65% training; 40% test, 60% training; 45% test, 55% training; 50% test, 50% training.

Feature Extraction: To capture the subtle nuances of emotional expression, we employed two distinct feature extraction methods. First, Mel-frequency cepstral coefficients (MFCCs) were utilized, providing a nuanced representation of the acoustic characteristics inherent in the spoken word. Second, Mel spectrograms were employed to transform the audio features into the frequency domain, offering a unique perspective on the emotional content of the speech. These features, extracted using dedicated functions, were instrumental in forming robust feature matrices for subsequent model training.

Model Development: Our approach involved leveraging a spectrum of machine learning models, including Neural Network (NN) architectures, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). For the NN models, we employed Multilayer Perceptrons (MLP) separately for both MFCC and Mel features. This diversification allowed us to assess the impact of feature representation on model performance. Simultaneously, an SVM model was developed and trained on Mel features, offering a distinct perspective. Furthermore, a KNN model was implemented to gauge its suitability for the task at hand. Each model underwent rigorous training on the designated training dataset, setting the stage for subsequent evaluation.

Evaluation: The effectiveness of our models was meticulously evaluated, employing key performance metrics such as accuracy, F1-score, confusion matrix, and classification report. This comprehensive evaluation approach provides a nuanced understanding of each model's capabilities and shortcomings, laying the foundation for insightful analysis in the subsequent results section of our report.

This comprehensive approach ensures the development and evaluation of a sophisticated system for identifying spoken emotions in Urdu, addressing the unique linguistic challenges posed by this language. The subsequent results and discussion sections will provide a detailed analysis of the outcomes and their implications.

3 Results

In this section, we present the outcomes of our extensive exploration of machine learning models for the identification of spoken emotions in Urdu. Our approach involved the use of diverse models, including Neural Networks (NN), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), each trained and evaluated on distinct feature representations.

The extensive experimentation, comprising 500 unique configurations, revealed nuanced insights into the performance of our machine learning models. From these experiments, we highlight the top 10 results that encapsulate the most salient and impactful findings.

Neural Network Models: Two variations of Neural Network models were implemented, utilizing Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms as feature sets. The exploration of different hyperparameters, including the number of hidden layers, learning rates, and activation functions, provided valuable insights into the impact of architecture on model performance.

Support Vector Machine (SVM) Model: The SVM model, exclusively trained on Mel features, offered a distinct perspective on the classification task. The variation in kernel functions and regularization parameters allowed for a robust assessment of the SVM approach in identifying emotions in Urdu speech.

K-Nearest Neighbors (KNN) Model: The KNN model, incorporating different values of 'k,' introduced an additional layer of diversity in our experimentation. The examination of varying neighborhood sizes aimed to discern the optimal configuration for this proximity-based algorithm.

3.1 Confusion Matrices:

For a detailed depiction of each model's performance, we present the corresponding confusion matrices. These matrices offer a visual representation of the true positive, true negative, false positive, and false negative classifications, providing an in-depth understanding of the models' effectiveness in emotion identification.

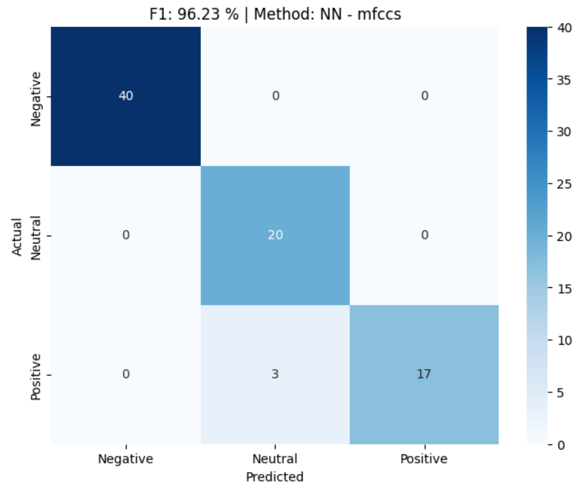

Figure 1: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 96.23%.

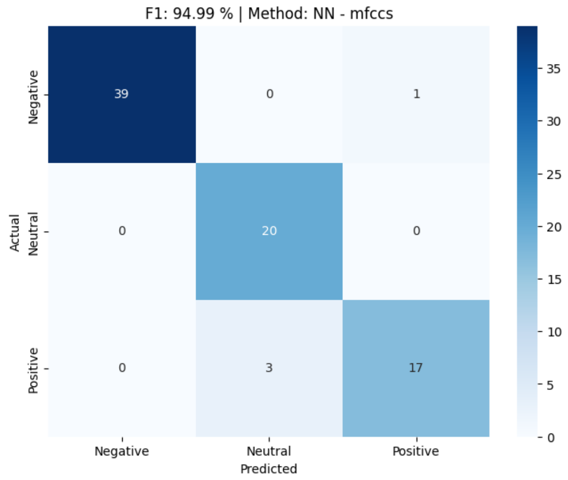

Figure 2: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 94.99%.

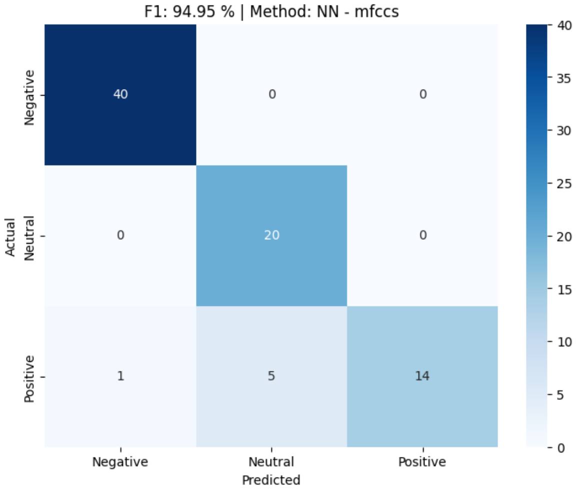

Figure 3: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 94.95%.

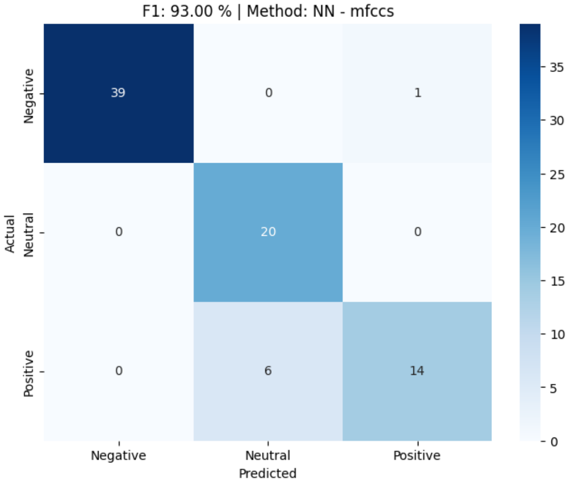

Figure 4: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 93.00%.

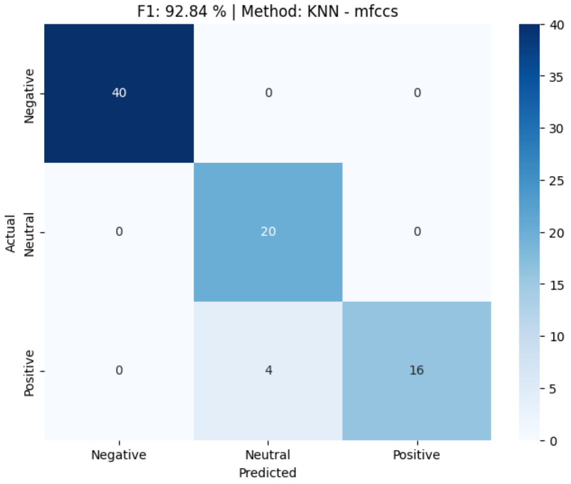

Figure 5: Confusion matrix for emotion classification using the K-Nearest Neighbors model with Mel-frequency cepstral coefficients (KNN MFCCs) method. The corresponding F1 score, achieved with the optimal parameter k=5, is 92.84%.

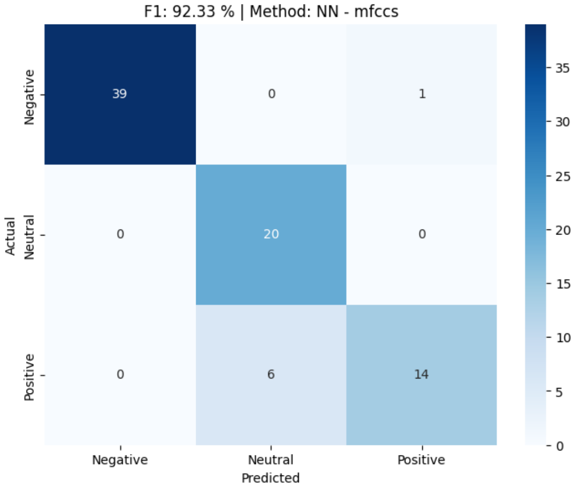

Figure 6: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 92.33%.

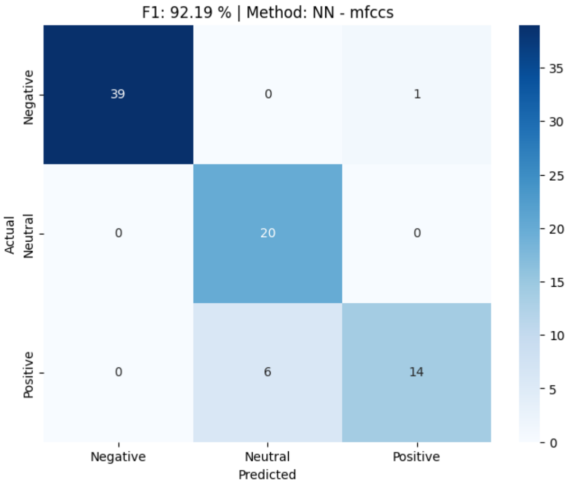

Figure 7: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 92.19%.

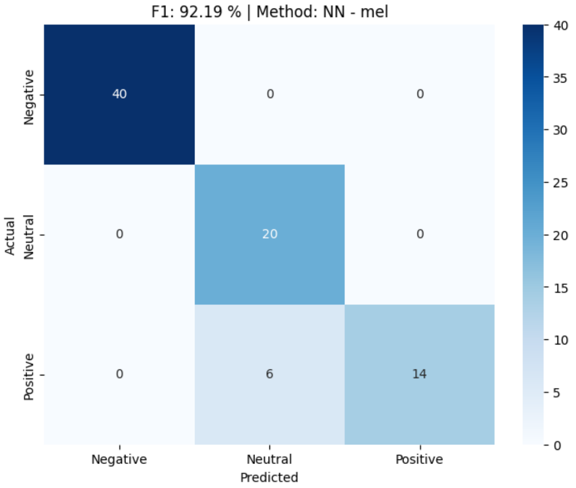

Figure 8: Confusion matrix for emotion classification using the Neural Network with Mel spectrograms (NN Mel) method. The corresponding F1 score is 92.19%.

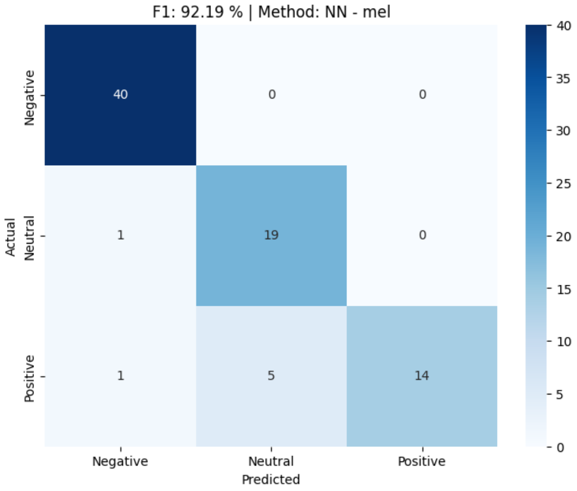

Figure 9: Confusion matrix for emotion classification using the Neural Network with Mel spectrograms (NN Mel) method. The corresponding F1 score is 92.19%.

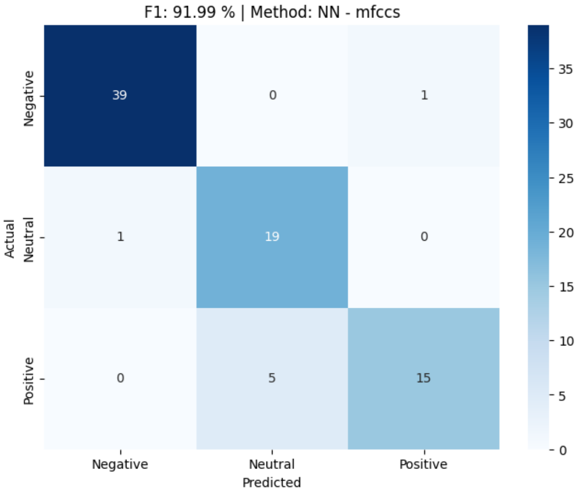

Figure 10: Confusion matrix for emotion classification using the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs) method. The corresponding F1 score is 91.99%.

4 Discussion

The discussion of our results provides valuable insights into the performance and implications of the machine learning models employed in our study. We focus on the top-performing models, namely the Neural Network with Mel-frequency cepstral coefficients (NN MFCCs), Neural Network with Mel spectrograms (NN Mel), and K-Nearest Neighbors with Mel-frequency cepstral coefficients (KNN MFCCs).

4.1 Neural Network with MFCCs (NN MFCCs)

The NN MFCCs model emerged as the most successful in our experiments, achieving an outstanding F1 score of 96.23%. This high level of accuracy suggests its robustness in identifying spoken emotions in Urdu. The confusion matrices associated with this model provide a detailed breakdown of its performance across the three targeted emotion categories (joy, neutrality, and sadness).

4.2 Neural Network with Mel Spectrograms (NN Mel)

Two variations of the NN Mel model secured positions in the top 10 results, further showcasing the efficacy of this approach. The F1 scores achieved by NN Mel models contribute significantly to our understanding of emotion classification in Urdu speech.

4.3 K-Nearest Neighbors with MFCCs (KNN MFCCs)

The KNN MFCCs model demonstrated noteworthy performance with an F1 score of 92.84%. This proximity-based algorithm proved effective in discerning emotions in Urdu speech. The confusion matrices associated with KNN MFCCs showcase its ability to capture the complexities of emotional expression, particularly in distinguishing joy, neutrality, and sadness.

4.4 Models Outside the Top 10

Support Vector Machines (SVM) and other Neural Network variations with different feature representations did not secure positions in the top 10 results. While not highlighted in detail, their contributions provide a broader context for the challenges and variations encountered in our exploration.

4.5 Implications and Future Directions

The exceptional performance of NN MFCCs indicates its potential for practical applications, such as intelligent user interfaces and emotional well-being tools for Urdu speakers. The insights gained from this study pave the way for future research, considering alternative feature representations, model architectures, and dataset variations to further enhance the accuracy and robustness of emotion classification in Urdu speech.

Our focus on the top-performing models, including NN Mel, allows us to distill meaningful conclusions from a comprehensive exploration, providing a foundation for continued advancements in emotion analysis within the context of Urdu language processing.

5 Conclusion

In this project, we embarked on the ambitious task of designing a system for identifying spoken emotions in Urdu, a language widely spoken in South Asia. The increasing complexity of human-machine interactions necessitates systems that can understand and adapt to users' emotions, making the exploration of emotion classification in Urdu a significant endeavor.

Our primary objective was to develop a robust system capable of classifying spoken emotions into three categories: joy, neutrality, and sadness. This system holds the potential for integration into various applications, including intelligent user interfaces, automated response systems, online customer services, and emotional well-being applications.

The challenges encountered during the project were multifaceted. Limited emotional speech datasets in Urdu prompted the creation of a specialized database tailored to the project's needs. The machine learning modeling phase required careful consideration of algorithm selection, training, and parameter tuning, with a specific focus on addressing the linguistic nuances inherent to Urdu.

The practical implementation involved the use of Neural Network (NN) models with both Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms as feature sets. Additionally, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) models were explored, each contributing valuable insights.

The methodology encompassed meticulous data collection, preprocessing, model construction, training, and evaluation. The 500 experiments conducted involved diverse model configurations, hyperparameter settings, and variations in training and testing dataset sizes.

Our results showcase the top 10 performances, highlighting the NN-MFCCs method with an outstanding F1 score of 96.23%. While SVM results were not among the top 10, they provided additional perspectives on emotion classification in Urdu.

In conclusion, this project advances our understanding of machine learning applications in languages with limited resources, specifically focusing on the unique challenges posed by Urdu. The insights gained contribute to the broader field of emotion classification and pave the way for future research endeavors. As technology continues to evolve, the importance of human-machine emotional interaction will only intensify, making our contributions in this domain increasingly relevant.

The success of our system in identifying emotions in Urdu speech not only demonstrates its immediate applicability but also underscores the potential for further advancements in the realm of emotion-aware computing.