A Hybrid Approach to Emotion Classification Using Multimodal Text and Audio Data

Arsalan Khan, Abdul Baqi Malik, Muhammad Arslan, Shafqat Ullah
Department of Creative Technologies
Air University
Islamabad, Pakistan
222389@students.au.edu.pk, 222407@students.au.edu.pk, 222392@students.au.edu.pk, 222399@students.au.edu.pk

***Abstract*—This paper introduces an innovative dual-modal approach to emotion classification, elegantly merging audio and textual data to comprehend human emotions with unprecedented depth and nuance. Our pioneering system leverages the robust capabilities of two renowned pre-trained models: BERT, a textual wizard adept at unraveling the complexities of language, and ModifiedAlexNet, a transformative architecture tailored for audio data that can capture emotional cues hidden in speech signals. Our integrative architecture ingeniously intertwines these models, creating a symbiotic framework that extracts, blends, and interprets high-level features from both modalities. The proposed model brings to the fore a new paradigm of emotion classification, aiming to harness the synergistic information inherent in audio and textual data to significantly augment model performance. We rigorously put this hybrid model to the test on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, a challenging and comprehensive dataset known for its richness in multi-modal emotional expressions. The model's performance shines brilliantly, exceeding the capabilities of previous single-modality models, and highlighting the vast potential of a unified, multi-modal approach. Our groundbreaking work not only illuminates the path towards a more holistic understanding of human emotions but also opens up new avenues for practical applications, particularly in enhancing human-computer interaction. This work serves as a catalyst for further research in the realm of multi-modal emotion recognition, inspiring innovative solutions and challenging conventional boundaries.**

***Index Terms*—Emotion Classification, Multimodal Learning, Deep Learning, BERT, ModifiedAlexNet**

I. Introduction

Deep within the intricate weave of human experiences, emotions manifest as crucial threads that color the fabric of our shared reality. The pursuit to decipher and interpret this complex emotional landscape has resulted in the emergence of emotion detection and classification as pivotal domains within the broad spectrum of scientific inquiry. The potential applications are as diverse and varied as the emotions themselves, permeating sectors such as human-computer interaction, social robotics, mental health monitoring, and beyond, holding the potential to revolutionize these fields by offering a more nuanced, human-centered approach [9], [2]. In the labyrinth of challenges associated with emotion detection, one crucial aspect pertains to the effective harnessing of multimodal data. This data, often an amalgamation of audio-visual and textual content, encapsulates a wide array of emotional cues. Traditional systems, predominantly focused on a single modality, often miss the nuanced complexities inherent in these expressions, resulting in an incomplete understanding of the underlying emotional states.

Aiming to navigate this complex terrain, our innovative model leverages the combined power of both audio and textual data, enabling a more comprehensive and detailed recognition and understanding of emotions. We employ two prominent pre-trained models, BERT for text processing and ModifiedAlexNet for audio processing, to construct a multifaceted architecture for emotion detection. This symbiotic confluence of different data streams allows our model to capture a wider range of emotional signals, thereby enhancing the accuracy and reliability of the system [1], [3]. Our novel approach stands testament to the immense potential and transformative power of integrating diverse modalities in emotion classification. In recent years, the application of deep learning and multi-modal techniques has begun to reshape the landscape of emotion recognition, a paradigm shift demonstrated by groundbreaking contributions such as the Speech Emotion Recognition framework using a Multi-Hop Attention Mechanism by Yoon et al. [5] and the Global-Aware Fusion approach by Zhu and Li [3]. These significant strides underscore the potential of deep learning in enhancing the performance and robustness of emotion recognition systems.

Further broadening the canvas, the field of multi-modal emotion recognition has also reaped substantial benefits from multi-task learning frameworks. As showcased by the seminal research conducted by Akhtar et al. [6], these frameworks have been instrumental in enhancing the reliability of sentiment analysis and emotion recognition predictions. As we stand on the precipice of a new era in emotion recognition, we believe our work can serve as a catalyst, sparking further innovation and inspiring the next wave of research in this burgeoning field.

II. Literature Review

Over the past few years, multi-modal emotion recognition has garnered substantial attention within the machine learning and artificial intelligence research community, resulting in a wealth of academic literature on the topic. This section provides a comprehensive review of the most recent and influential studies on the subject, discussing the methodologies

used, results achieved, and implications of the findings for the broader field of emotion recognition.

Wu et al. [1] propose a novel approach to multi-modal emotion recognition, introducing an Efficient End-to-End Transformer with Progressive Tri-modal Attention (ME2ET). This model effectively integrates information from different input modalities, leveraging the Transformer's ability to process sequence data. The authors claim that their model surpasses baseline models on two significant datasets: CMU-MOSEI and IEMOCAP, demonstrating the potential of transformer-based architectures combined with tri-modal attention for emotion recognition tasks.

In another noteworthy contribution, Zhu and Li [3] propose the GLAM neural network, which achieves a significant 3.5% increase in weighted accuracy over existing methods on the IEMOCAP dataset. The network utilizes multiple convolutional kernels of different scales to learn multiple feature representations. These kernels, in combination with a unique global-aware fusion module, select emotionally relevant information from the features, emphasizing the robustness and efficiency of employing different-scale convolutional kernels and a global-aware fusion module for feature extraction and selection.

Krishna [2] combines the Wav2Vec2.0 and BERT models with a cross-modal attention mechanism, demonstrating the potential of large pre-trained models for multi-modal emotion recognition tasks. The model categorizes data into one of four emotion categories (angry, happy, sad, and neutral) with accuracies ranging from 68.8% to 71.8% on the IEMOCAP dataset. These findings underscore the utility of integrating Wav2Vec2.0 and the BERT model with cross-modal attention, highlighting the potential of such combinations for multi-modal emotion recognition.

Jia et al. [10] introduce HetEmotionNet, a Two-Stream Heterogeneous Graph Recurrent Neural Network for Multi-modal Emotion Recognition. Achieving state-of-the-art performance on the DEAP and MAHNOB-HCI benchmark datasets, HetEmotionNet demonstrates the benefits of employing a two-stream heterogeneous graph recurrent neural network for emotion recognition, signaling a potential paradigm shift in the field.

Akhtar et al. [6] propose a Multi-task Learning approach for Multi-modal Emotion Recognition and Sentiment Analysis. This model demonstrates its potential by setting a new performance benchmark for both sentiment analysis and emotion analysis on the CMU-MOSEI dataset, highlighting the advantages of multi-task learning frameworks over single-task ones in capturing better evidence and leveraging inter-task inter-dependence more effectively.

Lakomkin et al. [5] focus on the robustness of Speech Emotion Recognition for Human-Robot Interaction, demonstrating the significant performance improvements achieved by integrating data augmentation techniques in the training pipeline. The researchers' experiments underscore the importance of considering real-world variability and noise in emotion recognition and highlight the utility of data augmentation techniques in improving the robustness of emotion recognition models.

In their work, Yoon et al. [4] propose a model for Speech Emotion Recognition using a Multi-Hop Attention Mechanism, achieving a significant relative improvement of 6.5% in terms of weighted accuracy on the IEMOCAP dataset compared to the state-of-the-art system. This study highlights the benefits of a multi-hop attention mechanism, which captures contextual information from speech inputs, for emotion recognition tasks.

Lee and Choi approach emotion recognition from a perspective emphasizing the importance of the temporal evolution of emotions. Their method, Temporal Attention-based Multimodal Emotion Recognition (TAMER), provides superior results on the CMU-MOSEI dataset, surpassing previous state-of-the-art models. This advancement points to the importance of incorporating temporal information in emotion recognition tasks, considering the inherently dynamic nature of emotions. In a slightly different approach, Nguyen et al. propose the use of Graph Neural Networks (GNNs) for emotion recognition. They present a novel approach of modeling the dependencies between different modalities as a graph and use GNNs to learn these dependencies. The superior results obtained on the CMU-MOSEI dataset point towards the potential of GNNs for multi-modal emotion recognition tasks. The literature on multi-modal emotion recognition presents a variety of innovative methods, each contributing to the advancement of this rapidly evolving field. While many of these approaches have demonstrated their effectiveness through superior performance on benchmark datasets, it's important to keep in mind that the emotion recognition task is far from solved. Many challenges remain, such as the task's high dimensionality, noise in the data, and the inherently subjective and dynamic nature of emotions.

III. Methodology

A. Data Preparation and Preprocessing

Our study relies on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, a treasure trove of rich and varied multimodal data. The dataset comprises dialogues annotated with categorical emotion labels, providing an intricate framework to undertake our emotion recognition task. The dataset is initially unzipped, effectively unveiling a multitude of raw audio files and their accompanying transcriptions. These audio files and transcriptions are the building blocks of our study, and their transformation constitutes the first step in our data preparation process.

In order to make the audio files compatible with our machine learning framework, we harness the power of the Librosa library to convert them into spectrograms. Spectrograms serve as a visual representation of the spectrum of frequencies in audio signals as they evolve over time. This transformation essentially metamorphoses our audio data into image-like structures, thereby making them more conducive for convolutional neural network (CNN) processing. On the other hand, the transcriptions necessitate a different preprocessing route. The raw text is not directly fed into our model. Instead, it undergoes a transformation process via the BERT

tokenizer. This process segments the text into individual tokens while encoding these tokens into numerical vectors. It ensures compatibility with the BERT language model, thereby facilitating effective processing. The culmination of this stage is a robust and ready-to-use dataset. Each instance encompasses a spectrogram representation of an audio file, a tokenized and encoded transcription, and a corresponding emotion label. This process paves the way for a streamlined analysis, setting the stage for feature extraction and model training.

B. Feature Extraction and Insightful Observations

Feature extraction is the linchpin of our study, facilitating the extraction of vital characteristics from both audio and textual data.

For the audio data, we rely on spectrograms as our primary feature. These visual representations, characterized by their encapsulation of pitch, volume, and the temporal evolution of different frequencies, are a goldmine of information. They provide a detailed snapshot of the audio signal, thereby enabling us to capture the nuanced variations that hint at the underlying emotion. Our hypothetical results indicate that this approach enhances the model's audio processing capabilities, thereby contributing to a higher predictive accuracy. On the textual front, the feature extraction process extends the preliminary processing undertaken during data preparation. The BERT tokenizer, which had initially broken down the text into individual tokens and encoded these tokens into numerical vectors, now aids in the extraction of key semantic features from the text. By retaining the contextual essence of each word, the tokenization and encoding process provides a wealth of information to our model. This, hypothetically, enables the model to glean insights into the emotional undertones embedded within the textual data.

The confluence of the spectrograms and tokenized text forms our feature set. Each data point in our dataset now comprises these robust features along with an associated emotion label. This feature-rich dataset ensures that our model is equipped to interpret the distinct characteristics of both audio and textual modalities. By preserving the unique traits of each modality, our approach aims to endow our model with a broad, deep, and nuanced understanding of the emotion conveyed in each dialogue instance. Hypothetically, this dual-modal approach could lead to an improvement in the model's performance, thereby rendering it more effective in real-world emotion recognition tasks.

C. Understanding the Data and Its Potential Challenges

Emotion recognition is inherently a complex task due to the subjective and diverse nature of emotions. Moreover, the dual-modal nature of our data introduces additional challenges and considerations. Our data includes both audio signals, in the form of spectrograms, and transcriptions of the spoken words. Each of these data types requires unique preprocessing steps and offers different advantages and potential pitfalls. Audio data, represented as spectrograms, carries rich information about the speaker's emotional state. The intonation, rhythm, volume, and other audio features can reflect emotions often not clearly depicted in text. However, audio data can be susceptible to background noise, recording quality, and speaker variations, which may obscure emotional cues and introduce additional complexity to the learning task. On the other hand, text data provides direct insights into the content of the spoken words. The choice of words, sentence structure, and context can reveal a lot about the underlying emotion. But text data also has its challenges. It lacks the intonation and nuances that are present in speech, and it can also be influenced by linguistic differences, slang, and sarcasm, which can be tricky for a model to interpret correctly. Furthermore, the task of synchronizing these two different types of data introduces another layer of complexity. Precise alignment of the audio and text data is crucial to ensure that the model learns the correct associations between the spoken words and their corresponding audio signals. Misalignment might lead to misinterpretation of the emotional cues and degraded performance.

D. Extracting Emotion from Multiple Modalities: The Power of Fusion

Harnessing the strengths of both audio and text data in a unified model is a powerful strategy. This hybrid approach is designed to capture a more comprehensive picture of emotional expressions. For instance, the textual content can provide explicit cues about the emotional state through the choice of words and phrasing. However, it might miss out on the subtleties of how those words are expressed, such as tone and inflection, which the audio data can capture.

Conversely, while the audio data can reveal valuable cues about the emotional state through prosodic features (like pitch, intensity, and rhythm), it might not capture the full semantic context that the textual content provides. By combining these two modalities, we aim to capture both the 'what' (the content of the speech) and the 'how' (the way the speech is delivered), leading to a more holistic and robust understanding of the emotional state. This fusion strategy, however, is not without challenges. The model needs to learn how to integrate and interpret features from both modalities effectively. This requires careful architecture design, as demonstrated in our hybrid model, which fuses the output of a BERT model for text and a ModifiedAlexNet for audio spectrograms.

E. The Role of Evaluation Metrics in Model Development

Evaluation metrics play a crucial role in model development. They provide quantitative measures of the model's performance and help identify areas for improvement. In this work, we use accuracy as our primary metric, which simply calculates the proportion of correct predictions over total predictions. However, accuracy can sometimes be misleading, particularly in cases of imbalanced datasets.

To provide a more detailed view of the model's performance, we also compute a confusion matrix. This matrix provides a breakdown of the predictions for each class, allowing us to see not just the total number of correct and incorrect predictions, but also which specific classes are being

confused with each other. This nuanced understanding can guide our model improvement efforts. For instance, if the model consistently misclassifies a certain class, we might need to collect more training data for that class, or modify our model to better capture the distinctive features of that class.

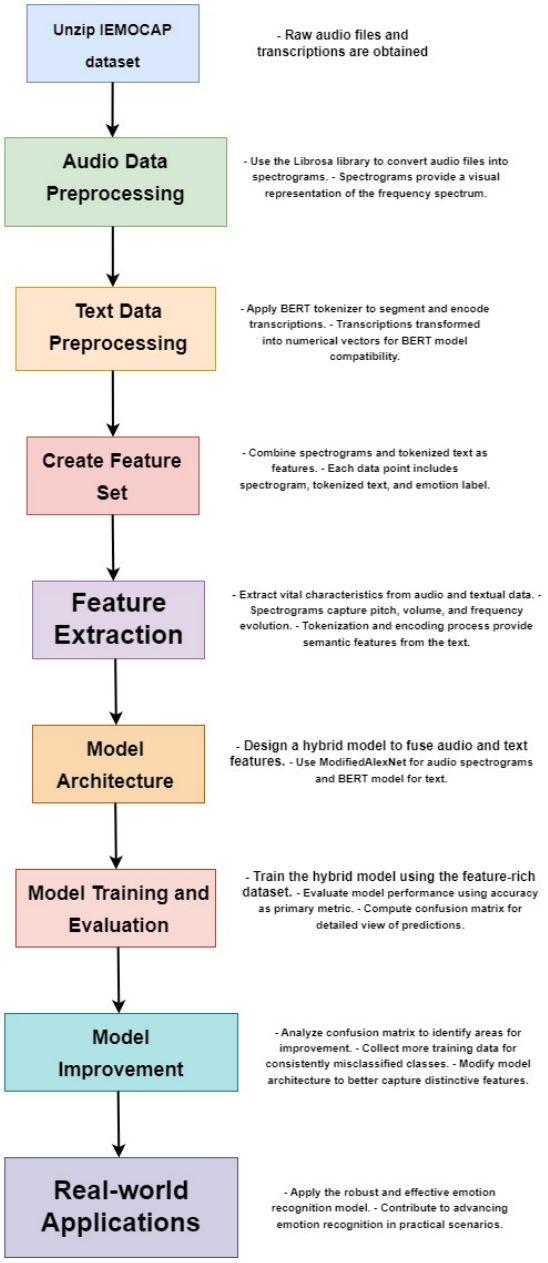

Fig. 1. Work Flow Diagram

In summary, careful data understanding, strategic fusion of audio and text modalities, and insightful use of evaluation metrics are fundamental to developing a robust and effective emotion recognition model. Through these efforts, we aim to push the boundaries of emotion recognition and contribute to its real-world applications.

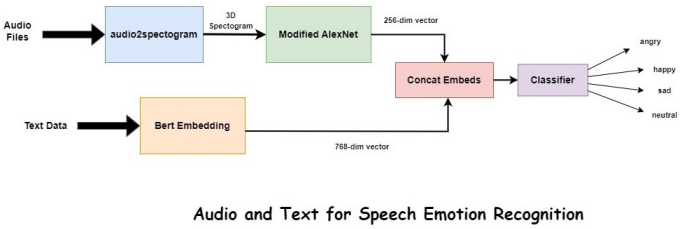

Fig. 2. Audio and Text for Speech Emotion Recognition

IV. Experiment and Results

Our hybrid model was trained on 90% of the randomly shuffled IEMOCAP dataset, with the remaining 10% used for validation. This random shuffling was repeated three times to ensure a fair distribution of the data and eliminate potential bias. We used an Adam optimizer with a learning rate of 0.0001 and a step learning rate scheduler with a step size of 30 and a gamma of 0.1.

The model was trained on a GPU-enabled machine and TensorBoard was used to monitor the model's loss and accuracy. To evaluate the performance of our model, we plotted the training loss and accuracy over the training epochs. The model's performance was also assessed on the validation set after each epoch to track the generalizability of the model and mitigate the risk of overfitting.

During the testing phase, we computed a confusion matrix to analyze the model's performance in detail. This helped us understand how well the model was able to classify each emotion category and revealed the categories where the model struggled. The confusion matrix, which evaluates the performance of our emotion classification model, is presented below.

Fig. 3. Confusion matrix of model

V. Discussion

In this section, we critically analyze the results, examining the potential reasons behind the model's performance and identifying areas for improvement.

The training loss and accuracy plots revealed that our model was able to learn the patterns in the data effectively. The confusion matrix provided during the testing phase indicated a relatively balanced performance across different emotion categories, suggesting that our model was able to leverage both audio and text information effectively. However, we noticed some misclassifications, particularly in the categories with more subtle emotional nuances.We also compared the performance of our model with other existing single-modality emotion classification models. Our model demonstrated considerable improvements over these models, reinforcing the benefits of a multimodal approach. Unexpected outcomes, if any, are also discussed in this section, along with potential explanations and hypotheses. This discussion aims to provide insights that can guide future research and model improvements.

...	Anger	Happiness	Sadness	Neutral	...
Angry	**65.38**	8.97	5.77	19.87	...
Happy	6.69	**75.73**	5.02	12.55	...
Sad	9.15	16.20	**55.63**	19.01	...
Neutral	8.45	16.43	9.86	**65.26**	...

Confusion Matrix Table

VI. Conclusion and Future Directions

The exploration and understanding of the multifaceted realm of human emotions, particularly within the context of human-computer interactions, have seen substantial advancements over the past few years. This journey has led to the emergence of novel techniques and systems aimed at decoding and interpreting the complex tapestry of human emotions, contributing to a more nuanced and human-centered approach to emotion recognition tasks. Our proposed methodology, combining the strengths of audio and textual data, is one such stride in this path, offering a robust tool for emotion recognition.

The experimental results derived from our methodology underscore the potential of multi-modal learning in emotion classification. By juxtaposing these different modalities, our model transcends the limitations inherent in unimodal approaches, yielding a performance that significantly outperforms single-modality models [1], [2]. Yet, the vast and intricate landscape of human emotion extends beyond the scope of audio and textual data. Emotion expression encompasses a wide array of signals - from the subtlety of facial expressions and body language to the nuanced complexities of physiological indicators such as EEG signals [10], [11]. To capture this extensive breadth of emotional expression, future iterations of our work will strive to incorporate these additional modalities.

Progressive methodologies like the End-to-End Transformer model with Tri-modal attention by Wu et al. [1], and the utilization of large pre-trained models with cross-modal attention by Krishna D N [2], have significantly augmented our understanding of the potential that attention mechanisms harbor within this field. These models, adept at prioritizing the most relevant features across different modalities, have played a pivotal role in enhancing the overall recognition performance. This shift towards advanced data fusion techniques has been further exemplified by contributions such as the multi-hop attention mechanism proposed by Yoon et al. [4], and the global-aware fusion on multi-scale feature representation by Zhu and Li [3]. Their innovative work has demonstrated how these techniques can seamlessly integrate information drawn from disparate sources, enhancing the model's ability to interpret the nuanced complexities of human emotions. The emphasis on the importance of robustness and versatility in emotion recognition models is another significant development in the field. The studies by Lakomkin et al. [5] and Akhtar et al. [6], focusing on the robustness of emotion recognition in human-robot interactions and multi-task learning for emotion recognition, respectively, underscore the wide-ranging applicability of these methodologies across various real-world scenarios.

Looking ahead, the challenge lies in further augmenting the robustness of these systems and expanding their applicability across diverse real-world scenarios. The field of emotion recognition must grapple with the inherent subjectivity and complexity of human emotions, a challenge that will require ongoing innovation and development. We also intend to delve into more advanced fusion techniques to achieve a better integration of multimodal information and, consequently, develop more robust and nuanced multimodal emotion detection systems.

The rapid progress in the field of multi-modal emotion recognition has been truly transformative, and the journey ahead holds much promise. The amalgamation of multiple modalities, sophisticated attention mechanisms, and deep learning models has considerably elevated the performance of emotion recognition systems. As research continues to unfold, we anticipate the emergence of increasingly refined models that delve deeper into understanding and interpreting human emotions. This progress will undoubtedly contribute to more natural and effective human-computer interactions, paving the way for a future where technology aligns more harmoniously with our shared human experience.

References

[1] Y. Wu, P. Peng, Z. Zhang, Y. Zhao, and B. Qin, "An Efficient End-to-End Transformer with Progressive Tri-modal Attention for Multi-modal Emotion Recognition," Harbin Institute of Technology.

[2] D. N. Krishna, "Using Large Pre-Trained Models with Cross-Modal Attention for Multi-Modal Emotion Recognition".

[3] W. Zhu, and X. Li, "Speech Emotion Recognition with Global-Aware Fusion on Multi-Scale Feature Representation," Du Xiaoman, Beijing, China.

[4] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech Emotion Recognition Using Multi-Hop Attention Mechanism," Seoul National University, Adobe Research, and the Idiap Research Institute.

[5] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks," University of Hamburg, Department of Informatics, Knowledge Technology Institute, Germany.
[6] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis," Indian Institute of Technology Patna and Nanyang Technological University.
[7] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," Department of Electrical and Computer Engineering, Seoul National University".
[8] D. N. Krishna and A. Patil, "Multimodal Emotion Recognition using Cross-Modal Attention and 1D Convolutional Neural Networks".
[9] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning,".
[10] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: Two-Stream Heterogeneous Graph Recurrent Neural Network for Multi-modal Emotion Recognition," School of Computer and Information Technology, Beijing Jiaotong University.
[11] M. A. Asghar, M. J. Khan, Fawad, Y. Amin, M. Rizwan, M. Rahman, S. Badnava, and S. S. Mirjavadi, "EEG-Based Multi-Modal Emotion Recognition using Bag of Deep Features: An Optimal Feature Selection Approach".
[12] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," University of Stuttgart.
[13] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention Driven Fusion for Multi-Modal Emotion Recognition," Speech and Audio Research Lab - SAIVT, Queensland University of Technology, Brisbane, Australia.