**A Project on**

**Intelligent Model Design for Emotion Recognition from text**

***Submitted in partial fulfillment of the requirement for the award of the degree of***

Master Of Computer Application /

Master of technology

****

**Under The Supervision of Dr. Arun Kumar**

**Assistant Professor**

Submitted By

Vidisha Arvind(E23MTCG0004)

Kartikey Saini(E23MCAG0127)

**SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY**

**BENNETT UNIVERSITY, GREATER NOIDA, U.P**

# A blue and red text on a white background Description automatically generated

# School of Computer Science Engineering and Technology - Bennett University

**CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the project, entitled **“ Intelligent Model Design for Emotion Recognition ”** in partial fulfillment of the requirements for the award of the **Masters of Computer Application/Master of Technology** submitted in the School of Computing Science and Engineering of **Bennett University, Greater Noida**, is an original work carried out during the period of Jan 2024 to May 2024, under the supervision of **Dr. Arun Kumar**, Associate Professor School of Computer Science Engineering and Technology, Bennett University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Vidisha Arvind

Kartikey Saini

**CERTIFICATE**

The Project Intelligent Model for Emotion Recognition of Data Mining has been held on 1st May 2024 and his/her work is recommended for the award of Master of Computer Applications/Master of Technology.

**Signature of Examiner(s) Signature of Supervisor(s)**

**Signature of Program Chair Signature of Dean**

Date: May, 2024 Place: Greater Noida

# Abstract (250 word)

In recent years, emotion recognition has become an increasingly important research area, with applications ranging from mental health support to customer service and entertainment. In this study, we conducted a comparative analysis of transformer-based models for emotion recognition using three distinct datasets: MELD, GoEmotions, and DailyDialog. Our experimental results demonstrated the potential of transformer-based models in emotion recognition, with the RoBERTa model exhibiting notable performance on the MELD dataset and significant improvement after optimizing the threshold for each label on the GoEmotions dataset. However, the performance varied significantly across different emotions, indicating the need for further improvement. The BERT model provided valuable insights into the model's performance on core emotions, showing potential but also revealing areas for enhancement. Our study contributes to the development of robust and effective models for emotion recognition, with the potential to significantly impact various applications. The findings highlight the importance of careful data preprocessing and the selection of appropriate models for emotion recognition tasks. We achieved an accuracy of 0.83 on the DailyDialog dataset, outperforming previous studies.

**Keywords**

 Emotion detection, Transformers, BERT, Fine-tuning, Contextualized embeddings

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **Title** |  | **Page No.** |
| **Abstract** |  | **I** |
| **List of Table** |  | **II** |
| **List of figures** |  | **III** |
| **Chapter 1** | **Introduction**   * 1. Motivation   2. Problem statement   3. Aim and objectives of the study | **1** |
| **Chapter 2** | **Literature Survey**   * 1. A review of relevant literature and existing work related to the project.(with table) |  |
| **Chapter 3** | **Methodology**   * 1. Hardware/ Software Requirement   2. Proposed Methodology   3. Details on data collection methods and tools   4. Procedures and techniques used in the project |  |
| **Chapter 4** | **Result and Discussion**   * 1. Presentation of the findings or results of the project   2. Use of tables, charts, graphs, or other visuals for clarity. Explanation of result |  |
| **Chapter 5** | **Conclusion and Future Work**  **REFERENCES** |  |

1. **Introduction**

In recent years, there has been a growing interest in developing intelligent systems capable of recognizing and understanding human emotions. Emotion recognition technology holds immense potential across various domains, including healthcare, education, customer service, and human-computer interaction. By enabling machines to interpret and respond to human emotions accurately, these systems can enhance user experience, improve communication, and even assist in mental health monitoring and support.

One of the key modalities for emotion recognition is text-based analysis, which involves extracting emotional cues from textual data such as social media posts, customer reviews, and chat conversations. Traditional approaches to text-based emotion recognition often rely on machine learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). For instance, studies have shown promising results in utilizing CNN architectures like mini-Xception to achieve high accuracy in emotion detection tasks. However, these traditional models may suffer from limitations such as limited generalization, especially when applied to diverse datasets or real-world scenarios [1].

Another emerging approach in text-based emotion recognition involves the utilization of advanced language models like BERT (Bidirectional Encoder Representations from Transformers). BERT-based models have demonstrated impressive performance in various natural language processing tasks, including sentiment analysis and emotion detection. By leveraging contextual information and semantic understanding, BERT-based models can capture nuanced emotions expressed in textual data. However, these models may face challenges such as fixed input length constraints, which can limit their effectiveness in processing longer texts or capturing complex emotional nuances [3].

Considerable attention has been devoted to automating the recognition of emotions, a fundamental aspect of human existence that significantly influences decision-making and interpersonal communication. While traditional approaches have primarily focused on identifying emotions through bodily behaviors such as hand trembling and voice pitch, there has been a growing interest in leveraging textual cues for emotion detection. However, detecting emotions in text presents significant challenges due to the inherent ambiguities and evolving linguistic expressions prevalent in everyday communication. Moreover, contemporary research suggests that emotion recognition extends beyond merely identifying basic psychological states like happiness, sadness, and anger.

In this research, we aim to develop an intelligent model for emotion recognition that integrates both text and image-based modalities. By combining traditional approaches with advanced techniques like BERT for text analysis and pre-existing models like Inception-V3 for image analysis, we seek to overcome the limitations associated with each modality and create a robust emotion recognition system. Our goal is to contribute to the advancement of emotion recognition technology and facilitate its integration into various real-world applications.

The remainder of this report is organized as follows. Section 2 reviews the relevant literature. Section 3 discusses the pre-processing of the research data, along with the research method, research model and performance index framework. Section 4 presents the experimental analysis and design. Section 5 discusses the research results, and conclusions are drawn in Section 6.

1. **Related Work**

Deep learning approaches have fueled breakthroughs in the domain of emotive computing and human-computer interaction, where emotion detection plays a critical role. Sowmya et al. [1] presented mini-Xception, a deep learning method designed for real-time emotion recognition from facial landmarks. They achieved remarkable accuracy rates by using convolutional neural networks surpassing Google Net, Caffenet, VGG16, and ResNet50. Similarly, Meena et al. [2] showed that CNNs are useful for sentiment analysis on images, especially when using the Inception-V3 transfer learning technique. These methods serve as the cornerstone of our investigation into various CNN architectures for the analysis of facial expressions.

Acheampong et al. [3] reviewed BERT-based techniques for text-based emotion recognition and demonstrated their competitive accuracy in identifying emotions from textual data. This is the foundation for our application of BERT variants to text based emotional content analysis, but the model was constrained to monolingual classification and could only work with a fixed number of tokens. Additionally, Kumar and Geetha [4] showed the promise of deep network approaches for textual emotion identification by combining natural language processing methods with recurrent neural networks (RNNs) but had less emotion labels due to consideration of labels from hashtags of the social media data.

Furthermore, Pan et al. [9] presented a multimodal method that included speech, EEG data, and facial expressions to recognize emotions with good accuracy rates and achieved an average accuracy of 98.27% for the improved Ghost Net model, surpassing the traditional Ghost Net model of 2.28 mb. Additionally, Zhihua et al. [10] did the multi-modal fusion approach for emotion categorization by combining text, audio, and video data that enhances the modeling of short and long-term dynamic correlation information between different modalities, improving emotion classification accuracy. Their approach seamlessly aligns with the core objective of our study, which is the integration of multiple data modalities to conduct thorough and comprehensive emotion analysis.

Beniwal et al. [6] introduced a hybrid BERT-CNN approach adept at handling social media text data, achieving impressive accuracy rates of up to 99%. Similarly, [7] proposed a few-shot ChatGPT system tailored for SMS, generating contextually relevant responses. Government Consultation Answering Language Model (GCALLM) by [8], utilizes transformers for seamless interactions between citizens and governmental services, stressing upon the need for continuous refining to sustain its effectiveness and studies such as [12] and [15] have shown custom CNN models optimized for resource constrained settings, exhibiting competitive accuracy rates in challenges involving the identification of emotions across different datasets.

Recent advances in model designs provide insights into using vision transformer architectures for image-based emotion detection. One example of this is the TransFER model, presented by Xue et al. [11], which combines attention processes and Vision Transformers for face expression identification. Comparably, Feng et al fine-tuned swin transformer [14], which blends optimization methods like Multiple Weights Optimality for similar emotion like surprise and fearful. [13] did dual domain affect fusion approach, showcasing improvements in accuracy rates by up to 14.9% through label fusion techniques. Furthermore, [16] laid the groundwork for leveraging Vision Transformers (ViT) in multimodal emotion analysis, albeit with accuracy rates of 50.05%. All these initiatives highlight the progress that has been achieved in developing emotion evaluation techniques, from hybrid model architectures to minimal neural network models that are specific to a certain domain.

One common limitation is the use of predetermined datasets, which may not capture the nuanced expressions present in real world circumstances. Furthermore, the issue of generalizing models learned on specific datasets to larger settings remains, necessitating the development of more adaptive methods. Furthermore, the interpretability of deep learning models, particularly those that use complex architectures such as transformers, must be improved to foster assurance and facilitate real-world use. Also, the computational needs of some approaches, such as vision transformers, impede scalability and real-time implementation, emphasizing the significance of optimizing and improving efficiency.

Table1. Summarization of Literature Review

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **References** | **Techniques** | **Database** | **Performance Metrics** | **Limitations** |
| [1] | Deep learning (mini-Xception) | FER2013.csv | Accuracy: 95.60% | Limited generalization, dependency on specific datasets |
| [2] | Transfer learning (Inception-V3) | CK+, FER2013, JAFFE | Accuracy: 99.57% (CK+), 73.09% (FER2013) | Reliance on pre-existing datasets, performance variance based on data quality |
| [3] | BERT-based approach | ISEAR, SemEval | Micro F1 scores: 0.688 (ISEAR), 0.695 (SemEval) | Fixed input length constraint, computational cost |
| [4] | Deep neural network and RNN | Twitter, websites, sentences, words, WhatsApp status lines | Accuracy: NRCLex (64.44%), Neural Network (99.56%), Simple RNN (83.36%) | Limited handling of varying text sizes, reliance on emoticons/hashtags for labeling emotions |
| [5] | CNN – 1D, LSTM | Emotion text data from kaggle. | Accuracy: 91% | Dependency on labeled training data, computational resources required |
| [6] | Hybrid BERT-CNN approach | Social media data from Instagram | Accuracy: 99% | Lack of detailed limitations provided |
| [7] | Few-shot ChatGPT system | SMS text messages | Accuracy: 0.7313 | Dependency on prompt design, potential output volatility |
| [8] | Transforms for generation | Governmental service consultation data | NA | Fine-tuning requirements, completeness of generated answers |
| [9] | Multimodal approach (facial, speech, EEG) | EMO-DB | Accuracy: 98.27% | Lower accuracy for fear expression, fusion method did not outperform individual modalities |
| [10] | Multi-modal fusion (text, audio, video) | MOSI dataset | Accuracy: 82% (Image+ Text) , 82.64% (Image + Speech + Text) | Computational complexity, need for extensive data augmentation |
| [11] | Multi-attention dropping, Vision Transformers | RAF-DB, AffectNet, FERPlus | Accuracy: 90.91% (RAF-DB), 66.23% (AffectNet), 90.83% (FERPlus) | Computational complexity, dependency on training data quality |
| [12] | EmotiEffNet (CNN-based) | AffectNet | F1-scores: 0.304 to 0.495 (expression recognition), 0.2862 to 0.4818 (action unit detection) | Generalization, robustness, real-time processing, multimodal integration |
| [13] | Dual domains affect fusion approach | AffectNet, MorphSet | Accuracy: 61.93% (AffectNet), 95.57% (MorphSet) | Subjective and noisy annotations, difficulty in categorizing emotions, under-exploration of dimensional models |
| [14] | Swin Transformer with multiple weight | RAF-DB, FERPlus, AffectNet | Accuracy: 90.38% (RAF-DB), 90.41% (FERPlus), 63.33% (AffectNet) | Difficulty in distinguishing similar expressions, potential for misclassification |
| [15] | Custom Lightweight CNN Model (CLCM) | Affectnet, RAF-DB, FER-2013 | Accuracy: 83% (RAF-DB), 54% (AffectNet) | Transfer learning limitations, challenges in recognizing complex emotions |
| [16] | Vision Transformers (ViT) | AffectNet | Accuracy: 50.05% | Limited handling of nuanced expressions, potential misclassification |

1. **Methodology**

**4.1 Database**

This study employs three distinct emotion classification datasets to evaluate and enhance the performance of our models. MELD, a multimodal conversational dataset from the TV series "Friends", presents the challenge of understanding contextual nuances and speaker dynamics. GoEmotions, a large-scale dataset of Reddit comments, tests the model's ability to handle a wide range of emotion categories. Lastly, DailyDialog, an international dataset of self-reported situations, assesses the model's capacity to understand cultural differences in emotion expression. By training and testing on these diverse datasets, we aim to develop robust and effective models for emotion recognition.

**4.2 Data Preprocessing**

The major focus was on  refining the datasets to enhance the performance of our emotion classification models. For the MELD dataset, we considered only two columns: the utterance, which contained the textual part, and the emotion, while discarding the sentiments and other remaining columns. This allowed us to concentrate solely on the emotion classification task.Similarly, for the GoEmotions dataset, we narrowed our focus to four core emotions instead of the original 28 emotions. This decision was made to improve the model's ability to distinguish and classify these fundamental emotions accurately.Similar process was also followed for the DailyDialog dataset of approximately 13k rows.

A diagram of a software development process

Description automatically generated

Fig.1 Overall Flow of the work

Prior to the modeling phase, we performed a series of text cleaning operations on all datasets. This involved removing special characters and emojis and converting all text to lowercase. Subsequently, we applied label encoding to assign a unique number to each emotion, facilitating the model's interpretation of the target variable. For tokenization, we utilized the BERT tokenizer, which is specifically designed to work with the BERT model. The tokenized sentences were then segmented and positionally tokenized, a process that involves adding special tokens to the beginning and end of each sentence and encoding the position of each token in the sentence. This step is crucial as it allows the BERT model to understand the context and order of the tokens, enabling it to generate more accurate and meaningful representations as shown in fig.1.

**4.3 Proposed Methodology**

**4.3.1 Multimodal multi-part dataset for emotion recognition (MELD)**

The four variants of the BERT model i.e. Roberta, ALBERT, Distil BERT, and BERT of base configuration, were employed for emotion detection in text data using the MELD dataset. The methodology for all four models was similar and included data preprocessing, model fine-tuning, and evaluation. The MELD dataset contains utterances with corresponding emotion labels, which were mapped to integers for model training and evaluation purposes. The tokenizers for each model were utilized to tokenize the utterances and create input sequences with a maximum length of 50 tokens. The tokenizers added special tokens and generated attention masks for the input sequences.

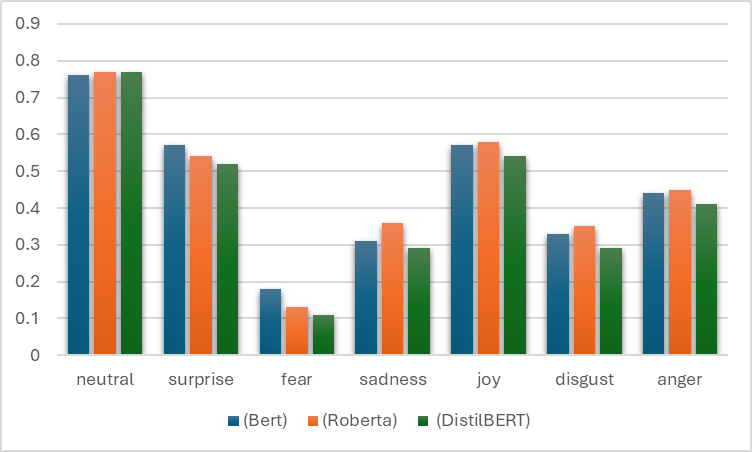


Fig.2 Overall accuracies achieved for different BERT methods

The dataset was then split into training, validation, and test sets. Models were initialized with their respective pre-trained configurations and were fine-tuned for sequence classification with seven output classes corresponding to the emotion labels. The models were trained on a GPU P100.The training process for all four models used a batch size of 16 and was run for 5 epochs. The AdamW optimizer was employed with a learning rate of 5e-5 and epsilon of 1e-8. A linear warmup scheduler was used with 10% of the total training steps as the warmup phase. The models were trained using a cross-entropy loss function in Fig.2.

**3.3.2 GoEmotions**

The two transformer-based models, RoBERTa and BERT, for multi-label text classification on the Go Emotions dataset. The pretrained roberta model from the Hugging Face model hub, which has been fine-tuned on this specific dataset. The model was used to predict the probabilities of each emotion label for every text instance in the test set. The hyperparameters used in this study were primarily the default ones provided by the Hugging Face library. However, we did not set a top-k value, allowing the model to return probabilities for all labels. To convert these probabilities into binary predictions, we introduced a threshold. Initially, we used a threshold of 0.5, meaning that any label with a predicted probability greater than 0.5 was considered positive depicted in Fig.3.

A graph of a bar chart

Description automatically generated with medium confidence

Fig.3 Label with a predicted probability greater than 0.5 threshold.

For the BERT model, the focus was on the core emotional labels, isolating six core emotions i.e. anger, fear, sadness, joy, surprise, and disgust from the original 28 categories. The dataset was structured using a custom dataset class, which facilitated the division of the data into training and testing sets. Input sequences were prepared with a maximum length of 64 tokens, ensuring uniformity in input dimensions. The labels were encoded into a dictionary mapping for model compatibility.

**4.3.3 DailyDialog**

The dialog dataset was divided into a training set of 7934 samples and a validation set of 3393 samples. The text data was preprocessed for use with BERT by setting the preprocess mode to 'bert'. The BERT model and vocabulary were automatically downloaded, and a maximum length of 350 was set to reduce memory usage and improve speed. The text was then preprocessed, and the train and test data were prepared.BERT-based text classifiers were trained using the preprocessed data. The model was trained using the One Cycle Policy with a maximum learning rate of 2e-05 for three epochs. The batch size was set to 6. The model's performance was evaluated using precision, recall, F1-score, and accuracy as achieved in Fig. 4.

A graph of different colored bars

Description automatically generated

Fig.4 Accurcaies obtained for DailyDialog for bert approaches VS LSTM.

1. **Experimental Results and Discussion**

**5.1 Multimodal multi-part dataset for emotion recognition (MELD)**

An empirical investigation employing four distinct transformer-based models RoBERTa, ALBERT, DistilBERT, and BERT was conducted, and each model underwent comprehensive evaluation, emphasizing test accuracy, validation F1 scores, and performance across diverse emotion classes. The RoBERTa model exhibited notable performance with a test accuracy of 0.6215 and achieved its best validation F1 score of 0.5894. This model demonstrated proficiency in classifying neutral, surprise, joy, and anger emotions, yet encountered challenges particularly with fear, sadness, and disgust emotions. Potential factors contributing to these disparities include class imbalance within the dataset and the nuanced complexity inherent in certain emotional categories.

Similarly, the ALBERT model attained a test F1 score of 0.6602 and a test accuracy of 0.5467, showcasing strengths in classifying neutral and joy emotions while encountering difficulties with surprise, fear, sadness, disgust, and anger emotions. Analogous to the RoBERTa model, the ALBERT model's performance is subject to class distribution disparities and the intricate nature of emotional classification tasks. Furthermore, the DistilBERT model achieved a test accuracy of 0.6061 and its highest validation F1 score of 0.6114, demonstrating proficiency in neutral, surprise, and joy emotion classification while struggling with fear, sadness, disgust, and anger emotions. The model's performance characteristics echo those of the RoBERTa and ALBERT models, suggesting common challenges related to emotion complexity and dataset balance.

Lastly, the BERT model yielded a test accuracy of 0.6176 and achieved its highest validation F1 score of 0.7906, exhibiting competence in neutral, surprise, and joy emotion detection while encountering challenges with fear, sadness, disgust, and anger emotions. The shared patterns across models underscore the impact of class imbalance and hyperparameter selection on model performance, necessitating nuanced strategies to enhance emotion detection capabilities in Table.2.

An empirical evaluation of the LSTM model was performed, confirming its efficacy in emotion categorization tasks. The LSTM model had variable F1 scores across assessments, ranging from 0.62 to 0.92. Notably, the model performed well with F1 values of 0.81, 0.82, and 0.86, suggesting its ability to properly capture the correlations between input sequences and emotional categories. However, there were certain situations when the model struggled, as seen by the F1 score of 0.63. Despite these variations, the LSTM model demonstrates its promise in emotion recognition tasks, showcasing its adaptability and ability to handle a wide range of emotional categories. Further modification and optimization of the model parameters may improve its performance across all emotion classes, assuring resilience and dependability in real-world applications.

Table.2 Comparison of accuracies achieved for existing models along with BERT approaches in MELD.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Emotions** | **F1(Bert)** | **F1(Roberta)** | **F1 (DistilBERT)** | **F1 (LSTM)** |
| neutral | 0.76 | 0.77 | 0.77 | 0.81 |
| surprise | 0.57 | 0.54 | 0.52 | 0.63 |
| fear | 0.18 | 0.13 | 0.11 | 0.82 |
| sadness | 0.31 | 0.36 | 0.29 | 0.92 |
| joy | 0.57 | 0.58 | 0.54 | 0.89 |
| disgust | 0.33 | 0.35 | 0.29 | 0.62 |
| anger | 0.44 | 0.45 | 0.41 | 0.86 |

**5.2 Go Emotions**

The RoBERTa model showed promising results, with the simple mean of labels increasing by 20.2% and the weighted average increasing by 10.4% after optimizing the threshold for each label. However, the model's performance still varies significantly across different emotions, suggesting that there is room for further improvement while the BERT model, allowed the focus on the core emotions and provided valuable insights into the model's performance on these specific categories. The results showed that the model performed well on certain emotions but struggled with others, indicating the need for further fine-tuning or data augmentation. The LSTM model performed well throughout assessments, with F1 values ranging from 0.89 to 0.95. This demonstrates its ability to appropriately distinguish numerous emotional subtleties. However, although the model excelled in certain emotion categories, it struggled in others, indicating possible areas for additional development. These findings demonstrate the model's performance in emotion classification tasks while also emphasizing the significance of ongoing refining to improve its capabilities across several emotional characteristics in Table.3.

Table.3 Comparison of accuracies achieved for existing models along with BERT approaches in Go Emotions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Emotions** | **F1(Bert)** | **F1(Roberta)** | **F1 (LSTM)** |
| anger | 0.92 | 0.51 | 0.91 |
| fear | 0.85 | 0.68 | 0.89 |
| sadness | 0.93 | 0.59 | 0.95 |
| joy | 0.89 | 0.63 | 0.93 |

**5.3 DailyDialog**

The trained BERT-based text classifier was evaluated on the validation set, and the results were reported in terms of precision, recall, F1-score, and accuracy. The model achieved an accuracy of 0.83, with a macro-average F1-score of 0.83 and a weighted-average F1-score of 0.83. The model performed well across all classes, with F1-scores ranging from 0.81 to 0.86. The model was also tested on a sample message, 'I’m not sure if I will get a job or not', and it correctly predicted the emotion as 'sadness' in Table.4.

The LSTM-based text classifier was rigorously evaluated on a validation set, with performance parameters such as precision, recall, F1-score, and accuracy carefully analyzed. The model achieved an accuracy of 0.73, with a macro-average F1-score of 0.72 and a weighted-average F1-score of 0.72, indicating its ability to capture the intricacies of textual material. The model performed admirably across several emotion classes, with F1-scores ranging from 0.71 to 0.74, demonstrating its ability to detect a wide range of emotions. Furthermore, to determine its practical applicability, the model was tested using a sample message, 'I'm not sure if I'll obtain a job or not', and effectively predicted the underlying emotion as ‘sadness'. These findings highlight the LSTM model's suitability for real-world applications, demonstrating its capacity to reliably categorize emotions in textual data, but with room for additional modification and optimization to improve performance across all emotional aspects.

|  |  |  |  |
| --- | --- | --- | --- |
| **Emotions** | **F1(Bert)** | **F1(Roberta)** | **F1 (LSTM)** |
| joy | 0.85 | 0.62 | 0.71 |
| sadness | 0.81 | 0.68 | 0.74 |
| fear | 0.87 | 0.66 | 0.71 |
| anger | 0.76 | 0.69 | 0.72 |
| neutral | 0.84 | 0.67 | 0.73 |

Table.4 Comparison of accuracies achieved for existing models along with BERT approaches in DailyDialog

1. **Conclusions**

In conclusion, this study evaluated the performance of transformer-based models for emotion recognition using three distinct datasets: MELD, Go Emotions, and Daily Dialog. Unlike traditional deep learning or machine learning algorithms , where we used word2vec embedding , the contextual understanding of text was not taken care of but BERT helped to overcome the issue. The RoBERTa model demonstrated promising results in handling multimodal conversational data from MELD and a wide range of emotion categories in Go Emotions. Meanwhile, the BERT model provided valuable insights into the classification of core emotions in both MELD and Go Emotions. However, the performance of the models varied significantly across different emotions, indicating the need for further improvement. The data preprocessing phase played a crucial role in refining the datasets and enhancing the performance of our emotion classification models. By focusing on the utterance and emotion columns in MELD and narrowing down to four core emotions in Go Emotions, we were able to concentrate solely on the emotion classification task. Text cleaning operations, label encoding, and tokenization using the BERT tokenizer were also essential steps in preparing the data for the models.

The experimental results demonstrated the potential of transformer-based models in emotion recognition, with the RoBERTa model exhibiting notable performance on the MELD dataset and significant improvement after optimizing the threshold for each label on the Go Emotions dataset. However, the performance varied significantly across different emotions, indicating the need for further improvement. The BERT model provided valuable insights into the model's performance on core emotions, showing potential but also revealing areas for enhancement. This study contributes to the development of robust and effective models for emotion recognition, with the potential to significantly impact various applications such as mental health support, customer service, and entertainment. The findings highlight the importance of careful data preprocessing and the selection of appropriate models for emotion recognition tasks.

1. **Future Scope**

Subsequent investigations may examine supplementary characteristics, class imbalance in the datasets, fine-tuning the hyperparameters of the transformer-based models. One promising direction is to incorporate contextual information into the model as emotion can be highly dependent on the context in which the emotion is expressed, and incorporating information such as the speaker's tone, the topic of conversation, or the social and cultural norms of the situation could significantly improve the accuracy of the models.

Another potential direction is to explore cross-cultural emotion recognition because most emotion datasets are biased towards Western cultures, and developing models that can accurately recognize emotions in non-Western cultures is an important area of research. This could involve collecting and annotating new datasets that are more representative of diverse cultures, as well as developing models that can handle cultural differences in emotional expression. Clustering and dimensionality reduction, can be used to identify patterns in the data without relying on labeled examples. This could be particularly useful for emotion recognition, where labeled data may be scarce or biased. Developing interpretable models is also an important direction for future research.

**References**

1. Sowmya, B., Alex, S. A., Kanavalli, A., Supreeth, S., Shruthi, G., & Rohith, S. (2024). Machine learning model for emotion detection and recognition using an enhanced Convolutional Neural Network. Journal of Integrated Science and Technology, 12(4), 786-786.

2. Meena, G., Mohbey, K. K., & Kumar, S. (2023). Sentiment analysis on images using convolutional neural networks-based Inception-V3 transfer learning approach. International journal of information management data insights, 3(1), 100174.

3. Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789-5829.

4. Geetha, A. (2024). Emotion Detection from Text using Natural Language Processing and Neural Networks. International Journal of Intelligent Systems and Applications in Engineering, 12(14s), 609-615.

5. Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. Frontiers in Psychology, 14, 1190326.

6. Beniwal, R., & Saraswat, P. (2024). A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data. The Computer Journal, bxae018.

7. Nedilko, A. (2023, July). Generative pretrained transformers for emotion detection in a code-switching setting. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (pp. 616-620).

8. Han, J., kang Lu, J., Xu, Y., You, J., & Wu, B. (2024). Intelligent Practices of Large Language Models in Digital Government Services. IEEE Access.

9. Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., & Wang, S. (2023). Multimodal emotion recognition based on facial expressions, speech, and EEG. IEEE Open Journal of Engineering in Medicine and Biology.

10. Radzi, N. H. M., & Hashim, H. (2024). Research on Emotion Classification Based on Multi-modal Fusion. Baghdad Science Journal, 21(2 (SI)), 0548-0548.

11. Xue, F., Wang, Q., & Guo, G. (2021). Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3601-3610).

12. Savchenko, A. V. (2023). EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5715-5723).

13. Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2852-2861).

14. Feng, H., Huang, W., Zhang, D., & Zhang, B. (2023). Fine-tuning swin transformer and multiple weights optimality-seeking for facial expression recognition. IEEE Access, 11, 9995-10003.

15. Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A., & Lanata, A. (2024). Facial Emotion Recognition (FER) through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets. IEEE Access.

16. Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: facial emotion recognition with vision transformers. Applied System Innovation, 5(4), 80.