Comparative Analysis of DistilBert, XLNET, RoBERTa & BigBird for Emotion Detection in Conversational Text

Arslan Ahmed[1], Barkha[2], Tushar Kanjwani[3], Shahzad Nasim[4]

Abstract— **Detection of Emotion from Conversational Data is the key to improving human-computer interactivity and under- standing how emotions work in digital communication. This paper investigates the role of four latest sophisticated models based on Transformer architecture includes DistilBERT, XLNet, RoBERTa and Bigbird in achieving high accuracy for emotion recognition from text-based dialogues. This research employs a comprehensive methodology of text cleaning, label encoding, model training on the diverse dataset collected from social media platforms and chat logs for six key emotion classes: anger, fear, joy, love, sadness, and surprise. The model was evaluated using measures of accuracy, precision, recall F1-score and confusion matrices. The findings indicated significant model performance variations, where RoBERTa achieved the highest validation accuracy while BigBird remained robust across different metrics particularly in identifying intricate emotional subtleties. The classification of 'surprise' is often misperceived with both joy and love across all models, which the analysis also flagged as a consistent challenge. This study highlights both advantages and limitations of these models, while offering new grounds for improved accuracy in affective computing. The findings will have a substantial impact on how to create more empathetic and effective AI-driven communication tools like customer service, mental health therapy, social media analysis.**

Keywords— **Emotion Detection, Conversational AI, Natural Language Processing (NLP), Transformer Models, Affective Computing, Textual Analysis, Machine Learning, Emotional Dynamics, Sentiment Analysis**

INTRODUCTION

Emotion recognition from text is becoming increasingly essential as conversational AI interfaces such as chatbots, virtual assistants, and automated customer-service agents are integrated into everyday tasks. Understanding the emotional state of users enables these systems to interact more nat- urally, empathetically, and effectively. Emotion Recognition in Conversation (ERC) is a specialized field within affective computing that seeks to extract discrete emotions such as anger, fear, joy, love, sadness, and surprise from conversational text, going well beyond simple sentiment polarity [1], [2]. Ac- curate emotion detection improves user satisfaction, supports therapeutic and educational applications, and enhances content moderation on social platforms.

[1-2-3] Sukkur IBA University
[4] The Begum Nusrat Bhutto Women University, Sukkur
Country: Pakistan
Email: shahzad.nasim@bnbwu.edu.pk

Traditional sentiment analysis, while useful, frequently lacks the granularity required in domains where identifying subtle emotional cues is critical. Early approaches LSTMs, CNNs, and RNNs could handle text classification but struggled with capturing long-term dependencies and nuanced context [1] [3]. The advent of Transformer-based models marked a paradigm shift: architectures like BERT, XLNet, RoBERTa, and BigBird rely on self-attention mechanisms to encode both syntax and context more effectively [4], [5], [6]. Although these models have achieved state-of-the-art results in a wide range of NLP tasks, their comparative effectiveness specifically for emotion detection in conversation remains underexplored.

A notable gap lies in evaluating models designed for different trade-offs: while DistilBERT is optimized for efficiency, XL-Net leverages permutation-based training to capture bidirec-tional context without masking, RoBERTa employs optimized pretraining routines, and BigBird uses sparse attention for long input sequences [5], [6] [7]. Previous comparisons [8] have shown that Transformer variants can outperform tradi- tional models in emotion classification. However, few stud- ies have examined how differences in attention mechanisms, encoding paradigms, and sequence handling affect emotion detectionespecially in dynamic conversational text. Emotion detection in conversations poses unique challenges. Emotions in dialogue may hinge on prior speaker turns, and lexically similar emotions like joy and love or surprise and fear can be easily confused [2], [8]. Moreover, emotional labels are often imbalanced, with some emotions underrepresented, reducing model sensitivity to less common classes like surprise [2]. BigBird's long-sequence capacity, for example, may allow better tracking of emotional progression in extended dialogues, but its effectiveness versus performance-intensive architectures remains a question.

Our study addresses these challenges by conducting a direct, head-to-head comparison of four Transformer architectures [9] DistilBERT, XLNet, RoBERTa, and BigBird for the task of emotion detection in conversational text. We begin with a diverse, Kaggle-sourced dataset containing social media dialogues and customer-service interactions, manually anno-tated across six emotional categories. This dataset covers a broad spectrum of conversational styles and ensures that sub-tle, context-dependent emotion variations are included. Each model is fine-tuned using a standardized pipeline, involving data cleaning, label encoding, tokenization [10], and tun- ing with uniform hyperparameters. We measure performance across multiple dimensions accuracy, precision, recall, F1- score, and confusion matrices highlighting both overall and per-class variations [11].

Our research aims to provide several key contributions. First, it offers a controlled empirical benchmark for mod- ern Transformer-based models, highlighting how attention structure, context capture, and sequence length capabilities translate to emotional understanding. Second, it presents architectural insights: for example, the comparative speed-efficiency of DistilBERT, XLNet's handling of syntactic complexity, RoBERTa's contextual robustness, and BigBird's performance on narrative dialogue. Third, it sheds light on consistent classification challenges particularly the misclassification of surprise offering suggestions for addressing label imbalance and ambiguous emotion boundaries.

This investigation has practical significance. Real-world conversational AI applications such as mental health chatbots, educational tutors, and online content moderation tools re- quire accurate emotion detection. By revealing how archi- tecture choice influences model behavior in these contexts, we inform system designers about potential trade-offs. For instance, BigBird may help track emotional progression across long dialogues but may struggle with efficiency in real-time applications where DistilBERT excels.

In addition, highlighting the limitations of existing Transformer models such as difficulty recognizing surprise motivates future research directions. Effective solutions might include targeted data augmentation for rare emotions, hierarchical emotion labeling, or hybrid approaches combining text with speech or visual cues [12], [13], [14]. Our study thus forms a methodological bridge: combining large-scale emotion-labeled conversational data with comparative model analysis to guide next-generation emotion-aware systems.

The structure of this paper proceeds as follows. Section II reviews relevant literature in emotion recognition and transformer-based NLP, exposing prior work and remaining gaps. Section III details our methodological approach, including dataset preparation, preprocessing, tokenizer choice, model training, and evaluation framework. Section IV reports experimental results across each model. Section V offers an interpretation and discussion of these findings, examining architectural strengths, limitations, and implications. Section VI concludes by summarizing contributions, outlining theoretical and practical implications, and proposing directions for future work Section VIII.

LITERATURE REVIEW

In this section, we provide a comprehensive literature re- view on emotion detection based on text, specifically we examine the evolution of Emotion text-based detection based on statistical models [15], Machine Learning Models, [16] Deep Learning Models [17], and furthermore application of transformer-based modes. Additionally, we also explore some existing transformer-based models like BERT, [18] Distil- BERT, RoBERTa (A Robustly Optimized BERT Pretraining Approach), XLNet: Generalized Autoregressive Pretraining for Language Understanding and BigBird, is a sparse-attention based transformer which extends Transformer based models. By combining the current study, we identify the research gap and propose further discussion to improve accuracy, effectiveness, and efficiency of Emotion detection text-based subtasks with the provision of transformer-based models [9].

A. Evolution of Emotion text-based detection

Early approaches to emotion text-based detection task as tweets text, chat conversation text, or email body text classifi- cation problem, where the overall sentiment of the entire text was or sentence was determined [19] However, this approach could not detect nuanced opinions about specific emotions or sensitivity mentioned in the text. whereas emotion text-based detection emerged as a kickback to this limitation, aiming to provide a more detailed understanding of emotions by associating it with dimensions or features.

Text-based emotion detection has become increasingly prevalent due to its applications in understanding end-user emotions over time, conversation, or post in a more unstructured manner. There are various studies of Emotion text-based detection, models are Machine-learning (ML) models [16], Deep-learning (DL) [20] and transformer-based pretrained models [21].

B. Machine Learning Techniques

Machine Learning (ML) is a key area of Artificial Intelli- gence (AI), focusing on the development of advances statistical models and algorithms [16]. ML allows computer enough intelligent to learn form given data and make decisions or predictions without any complex or hard programming instructions [22]. The application of ML techniques marked a significant advancement in Emotion detection from text. Supervised learning models, such as Logistic Regression, Random Forest and Support Vector Machine (SVM) [23], were employed for Emotion Detection from text. These models performance and effectiveness depended on feature engineering, involving the careful extraction of key text features such as n-grams, part-of-speech, and embedding techniques. However, the performance of these methods depends on features selection. However manual feature engineering causes for limitation in both time consuming and efficiency, so to overcome these limitations, researchers explored DL techniques for Emotion detection tasks just to enhanced efficiency and accuracy.

C. Neural Network Techniques

Powered by rapid advancements in neural network tech- niques, Deep Neural Networks (DNNs) have achieved visible improvement in various aspects and purposes. This approach has led Emotion Detection from text research to transform from feature-based techniques to auto feature extraction likely DNN Models. Deep Neural Network architectures, such as recurrent neural networks (RNNs) [24], convolutional neural networks (CNNs) [25], and transformers [21] revolutionized the field of Emotion detection from text. This study utilized the Emotions detection datasets from Go Emotions dataset [dataset paper reference], and Emotions from Tweet Emotions datasets for Emotion detection form text. Whereas, for the Emotion detection task, Tweet Emotions, and GoEmotions dataset were utilized from the past 2 years. In addition to that, an RNN-based model that involves long short-term memory (LSTM) [24] [26], and a gated recurrent unit (GRU). These models have shown enhanced performance in capturing contextual information and learning intricate patterns within textual data. Attention-based LSTM and Bi-LSTM models have been widely proposed for Emotion Detection [27]. The ability of neural networks to automatically learn features that contributed to their goal in particular given tasks. This quality impressive to shift from old

feature-based methods to new Neural network architectures which has significant achievements in recent years. This Neural network approach bring a very good perspective and efficiency in text-based Emotion Detection. Whereas we found some limitation in the models Like RNN and its flavors like GRU, LSTM, Bi-LSTM, firstly we have to train these models on certain dataset and then testing on another dataset are not performing good on other dataset that mean we have to train our models on huge data for more efficiency so, in recent years we found some transformer- based pretrained models pretrained which are already trained on a very large dataset [28].

D. Transformers Based Techniques

The basic observation behind transformers were a Sequence-to-sequence modeling for machine translation. Later study demonstrates that Transformer-based pre-trained models (PTMs) can produce novel outcomes on many kinds of tasks. Transformer PTMs have excellent performance in text-based Emotion detection and other natural processing tasks [18]. Much more efficient than earlier methods in the obtain of Longitudinal dependencies and contextualized information or data. Transformers quality of self-attention mechanism that obtain Longitudinal dependencies within a text until the neural networks which process the data sequentially whereas transformer process word parallel to obtain the contextual information [21]. As Figure 1 shows that encoder takes a series of input symbols like [x1, x2,. ,xn] is encoded by continuous series of symbols like [z1,z2,. . . ..,zn], once z is generated, the decoder starts decoding to create the continuous series of output sequence [y1,y2,. yn] [9].

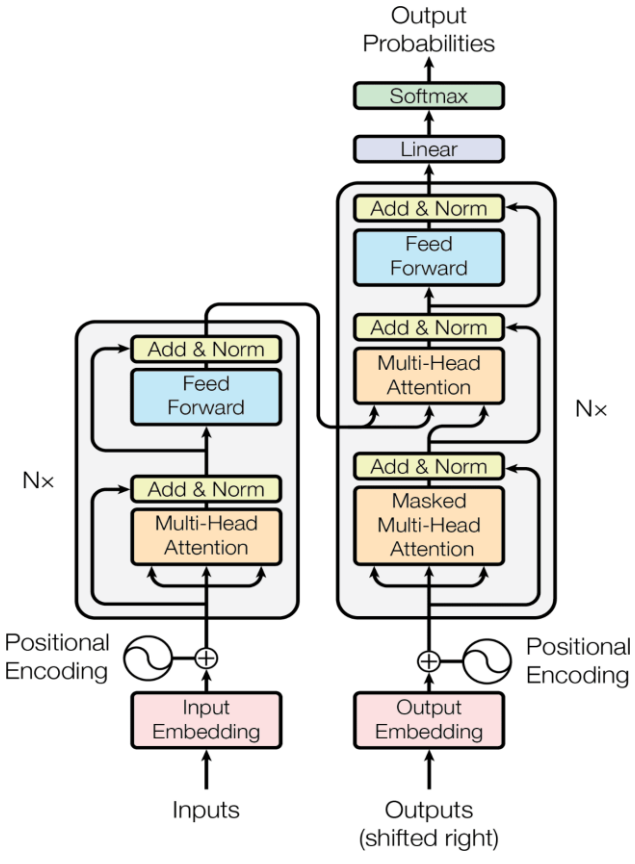

Fig. 1. Transformer model architecture

In the context ofText-based Emotion detection, encoder takes the conversation sentence and emotion as an input and encodes it into a highly detailed representation whereas the output side the decoder predicts the emotion according to their features. The self- attention mechanism of transformer records the features for better results [20]. Fine-tuning pre-trained transformer models for text-based emotion detection tasks became a common practice, allowing models to leverage large-scale pre-training on distinct textual data.

E. BERT

BERT (Transformers Bidirectional Encoder Representa- tions) has become more influential in recent years because of its powerful transformer-based model that has recast natural language processing tasks because of its bi-directional Transformer architecture context understanding becomes stronger [20] BERT performance across almost all domains of NLP such as Emotions detection in text, text generation, text summarization, text classification, sentiment analysis, question answering, and machine translation etc.

BERT's understanding of language in context has showed the noticeable efficiency in search engines, chatbots, and virtual assistants and getting more accurate responses and user enjoys their better experiences [28]. Although BERT-based models have also performed extraordinary in Text-based Emotion Detection tasks, including the need for domain-specific pre-training, handling multi labelled data for specific domains for more specific between two most similar Emotions Like Happy and Fun. Hence, to overcome these limitations, Pre- trained Models such as DistilBERT, RoBERTa, XLnet and BigBird [29] can be used for more accurate prediction text- based emotions detection tasks.

F. DistilBERT

DistilBERT is based on the concept of knowledge distil- lation introduced by [30]. It uses a distillation knowledge to train as a pre-trained model BERT's on very few features and in results good performance and accurate prediction [3]. This is same as BERT Transformer architecture uses pre-trained model on self-supervisor learning on textual data uses bi- directional encoder and decoder for contextual understanding of information while making prediction. Many researchers and developer prefer DistilBERT model because it's smaller in size and faster in training with any training loss perform better and more accurate [29].

G. RoBERTa

As in above sections we discuss some transformer-based architecture and modification to pre-trained models with their hyperparameters that how they improve their performance accordingly [3]. Further we add up the improvements by setting new models with its some hyperparameters to BERT approach and this configuration setup known as RoBERTa for Robustly optimized BERT approach. This model is trained using a masked language modeling goal and a next sentence prediction objective on a wide range of text data, including novels, web pages, and Wikipedia. [29]. The results RoBERTa can produce revolutionary results on emotion detection tasks, as shown by the researchers, and it depends on dataset nature too to find the best models fit on tasks.

H. XLNet

XLnet is another advanced algorithm of pre-trained language models. This algorithm is based on self-attention mechanisms, and its basic feature like BERT but little more different from BERT [31]. XLNet is an autoregressive modeling method called autoregressive self-attention. It is a combination of autoregressive (AR) and autoencoder (AE) [29]. This algorithm works like we want prediction of emotion XLnet fetch all the meaningful information from all possible nodes of language representations with better contextual understanding helps a lot in Emotions prediction for unseen data.

I. BigBird

As all above discussed Transformer-based models are more efficient deep learning models for NLP tasks. But limitation of these models are their attention mechanisms is much more dependent on sequence length as like memory size, batch size and so on which was big barrier by this barrier handled by a spare attention mechanism algorithm called BIGBIRD [7] introduced by Google research team.

It takes linear time as above models takes time quadratic dependent. It takes linear time, because of its spare attention mechanism as all other models works on full attention mechanism that's take time and effects the results too. BIGBIRD uses sparse attention on each layer of the input sequences X = [X1, X2, X3, X4,. Xn] [?]. For a directed graph D, the generalized attention with vertex set of V = [1,2, 3, n] Neighbors N(i), then the output vector of the generalized attention is given as:

$$Attn_D(X)_i = x_i + \sum_{i=1}^{Hd} \alpha(Q_n(x_i)K_n(X_{N(i)})^T).V_n(X_{N(i)})$$

where:
Qn and *Kn* are the query and key functions, respectively,
Vn is a value functions,
α is a scoring function, and
Hd respesents the head in numbers.
Also note *XN (i)* corresponds too the matrix formed by only stacking $\{xj : j \in N(i)\}$ and not all the inputs.

There are very few studies in the literature that explore the potential of BigBird and other sparse-attention-based transformers for emotion detection, explore the detailed comparative analysis from various domains, and the LLMs are not presented. However, in general, the main purpose of the study is to predict the different emotions of text conversation and also of tweets. Furthermore, this study does not present a comparative analysis with LLMs.

J. Comparison to Literature and Literature Gap

Although some research studies examined the model performance for NLP, general sentiment tasks or emotion recognition task, only very few made the comparison based on the emotion recognition in the conversational scenarios. For instance:

- In [8], applied different BERTs to generic emotion tasks but did not consider conversational nuances.
- In [12], presented speaker-aware RoBERTa, though they did not compare against sparse attention models like BigBird.

In addition, there are limited studies which address the under-representation of emotion labels (e.g., "surprise"), emotion label imbalance and context ambiguity (e.g., confusion be- tween joy and love). The majority of works ignores domain adaptation, trade-offs and real-time adoption.

There are, however, significant gaps in the current literature which restrict the progress towards building emotion detection systems that are robust to conversational text. Only a few works carry out thorough, head-to-head comparisons of Transformer models with the same experimental setting applied to dialog-based emotion recognition. Sparse-attention back-bone models such as BigBird also show promise for long-text processing but have been less extensively studied in this direction. Second, little attention has been paid to less frequent or inherently ambiguous emotion classes (e.g. "sur- prise"), which typically causes poor performance and neglects some nuances of emotion. Apart from model accuracy, little consideration has been paid to the balancing trade-offs among computational efficiency, inference latency, and deployment practicality. These limitations imply that there is space for broader studies that do not stop at benchmark comparisons between model types but which also account for real-time and resource limitations of actual models in the wild.

METHODOLOGY

A. Data Collection

The dataset employed in this research was obtained from Kaggle, a globally recognized platform hosting large-scale data science competitions and curated datasets for advanced analytical research. This particular dataset comprises a di- verse collection of conversational text samples extracted from multiple sources, including informal social media interac- tions, customer service exchanges, personal chats, and digital discussion forums. The richness of this dataset lies in its coverage of varied linguistic tones, grammatical structures, and conversational complexities, making it a highly suitable foundation for emotion detection tasks in natural language processing (NLP).

Each dialogue is treated as a data instance and stored in tabular format (rows x columns) where rows are the instance or records and 2 columns means the raw text or dialogue and its emotional label. This structure supports a completely smooth supervised learning and classification process. Crucially, such a dataset covers six universally recognized emotions for the comfort: anger, fear, joy, love, sadness, and surprise, consistent with existing emotion models in psychology and affective computing. Figure 2 shows the analysis of label distribution on the training images to emphasize the class balance and fairness in training and evaluation.

B. Annotation Method

The dataset was initially annotated for discrete emotional labels based on an improved schema defined by six core emotions: anger, fear, joy, love, sadness and surprise. Anno- tations: Following established guidelines in affective computing, annotations were done by more than one annotator to ensure

consistency and minimize subjective bias. We used its conversational dialogue and labeled each such dialogue with specific emotional labels making learning from it in a structured manner easy for training emotion detection models. This

methodical annotation enables a detailed analysis of the models' capabilities in recognizing and classifying nuanced emotional expressions.

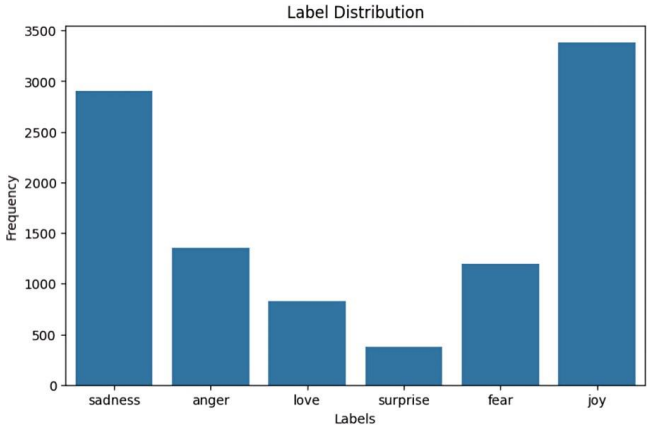

Fig. 2. Distribution of Emotional Labels in Training Data

C. Data Preprocessing

Preprocessing is a crucial stage in natural language comprehension, particularly when addressing fine-grained, semantic classification tasks like emotion detection [32]. The raw text was processed through a series of systematic transformations to minimize noise and unify the input across all models. These steps included:

- ***Lowercasing:*** All text was converted to lowercase to avoid case-sensitive mismatches.
- ***Punctuation Removal:*** All nonalphanumeric characters were eliminated using regular expressions to prevent irrelevant tokens from affecting the learning of the model.
- ***URL and Username Removal:*** The identifiers, links, and other metadata were removed to maintain the privacy and prevent leaking of sensitive information.
- ***Stopword Removal:*** All the ENGLISH (common) stop- words were eliminated, thus sharpening the focus of the model on emotionally charged words.
- ***Noise Reduction:*** Low-frequency tokens, emojis and escape characters were removed for readability.

The character distribution of printed sentence was also counted (see Figure 3), which was useful for learning the text complexity and the optimal token sequence length for training.

D. Label Encoding Techniques

Since the categorical emotion labels when used to train and predict required numerical inputs, Label Encoder was employed for efficient model training and predictions. Having every one of the emotions types a unique integer as mention in Table I, makes it much easier for our model to digest and learn from those texts. The encoding used was kept consistent through training, validation and testing datasets to ensure data fidelity which resulted in an accurate model. This encoding remained consistent throughout the training, validation, and testing phases, ensuring label integrity and reproducibility during evaluation and cross-model comparison.

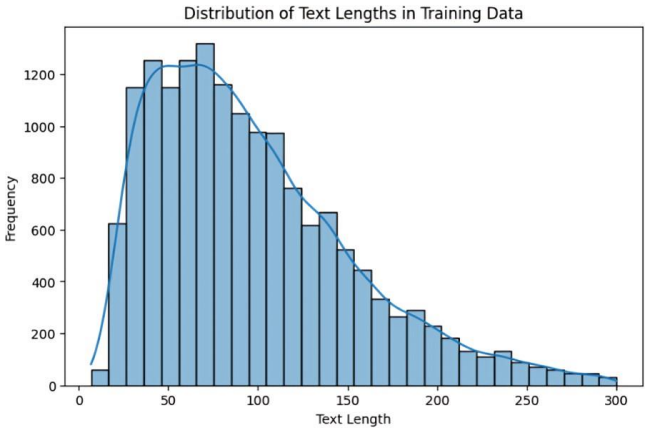

Fig. 3. Distribution of Text Lengths Across Conversational Samples

TABLE I
LABEL ENCODING SCHEME FOR EMOTION DETECTION

Label Mapping	
Category	**Label Encoding**
anger	0
fear	1
joy	2
love	3
sadness	4
surprise	5

E. Tokenization and Model-Specific Input Formatting

Tokenization was a crucial step in preparing the textual input for transformer-based models [10]. Each model architecture required its own tokenizer, capable of mapping the raw text to the token embeddings expected by the model:

- ***DistilBERT:*** Utilized the DistilBert Tokenizer, optimized for lightweight, efficient tokenization while preserving semantic fidelity.
- ***XLNet:*** Used permutation-enabled tokenization and encoding of positional permutations to model dependencies from both directions.
- ***RoBERTa:*** Employed a byte-level BPE tokenizer that is good for maintaining the structure and context of the sentence.
- ***BigBird:*** we utilized BigBird Tokenizer configured for long-seq processing and with padding/truncation settings that allowed maximum of 512 tokens.

Tokenization was performed with padding=True, truncation=True for obtaining a fixed-length input sequences that could be batch trained and adapted to the transformer model.

F. Model Selection Rationale

The models selected for this study *DistilBERT*, *XLNet*, *RoBERTa*, and *BigBird* were selected for their architectural richness and complementary capabilities of representing language semantics and emotions. Each model has contributed to a state-of-the-art invention of the Transformer family and targets certain inherent limitations in earlier designs:

1) DistilBERT: DistilBERT is a small, fast, cheap and light Transformer model based on BERT specifically pruned for fast inference. It utilizes knowledge distillation in the training phase to achieve strong efficiency, which is very suitable for

resource limited scenarios such as mobile devices or real-time systems [14].

In this study, it achieved good performance with minimal overfitting and quick convergence. While being lighter in terms of computational and memory requirements as shown working in model architecture shown in Figure: 4 and reduced in computational and memory cost. This model was selected because it effectively processes massive datasets with little loss of accuracy. It is efficient in understanding sparse conversational snippets. It is well suited for resource constrained scenarios like mobile and real time apps.

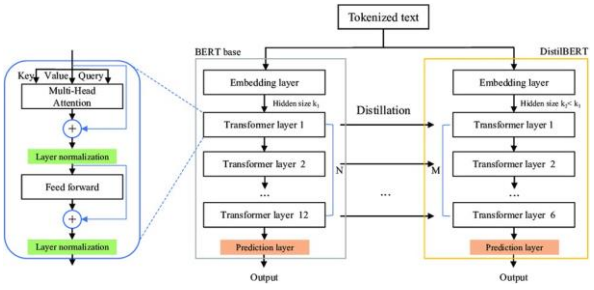

Fig. 4. DistilBERT Model Architecture

2) XLNet: XLNet improves on BERT by introducing a permutation-based training objective that captures bidirectional context without masking. Its autoregressive nature allows it to better model word dependencies, especially in longer or complex conversational sequences [33]. In this study, it demonstrated superior recall for emotions such as fear, which often depend on contextual nuance. This property makes XLNet good at dealing with long and complicated sequence structures as shown in Figure: 5 which is exactly what you want when analyzing conversations that include very complex emotional expressions.

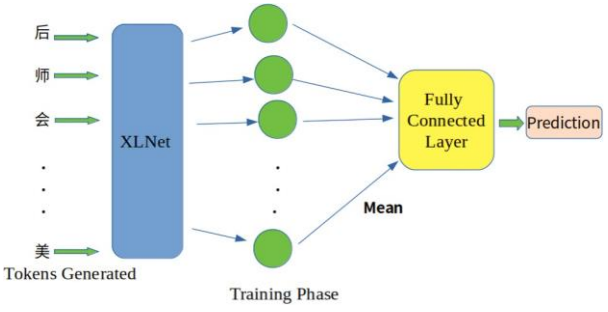

Fig. 5. XLNET Model Architecture

3) RoBERTa: RoBERTa refines the BERT architecture by optimizing the training regime: it uses dynamic masking, longer training durations, and larger batch sizes [6] [12]. It excels in understanding subtle emotional distinctions and proved to be the most balanced and accurate model in this research. It also showed strong generalization across classes and robustness in handling ambiguous or context-heavy phrases. Figure: 6.

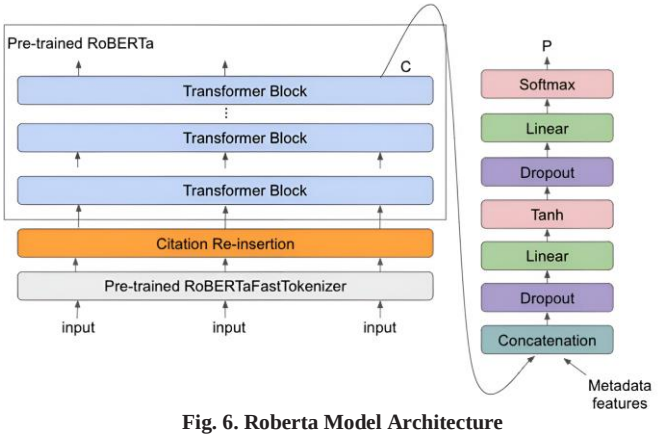

Fig. 6. Roberta Model Architecture

4) BigBird: BigBird is a Transformer based model that pro- vides wider context and long-range dependencies in constant num- ber of operations thus addressing some limitations of models such as BERT or RoBERTa, which capture only local context as of it's advance architecture Figure: 7. BigBird can handle much longer text sequences because it is using a type of sparsity attention such that each token attends to many fewer other tokens, which makes it ideal for analyzing very long conversations while preserving the relative information. It is useful for long threads of customer service interactions or Twitter conversations where emotional context can change multiple times during the conversation.

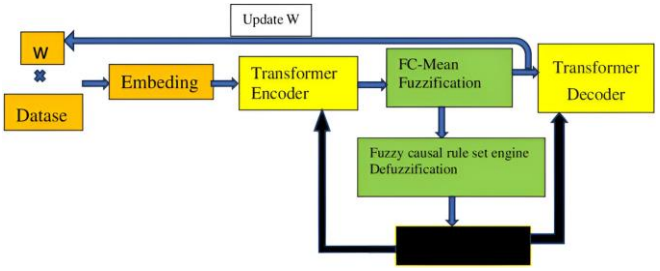

Fig. 7. BigBird Model Architecture

G. Model Implementation and Training

Each model was implemented using Hugging Face's transformers library and trained independently on the same training data. The training environments were standardized with shared hyperparameters (e.g., batch size, learning rate, epoch count) where applicable to allow fair comparison .

- Frameworks: TensorFlow for DistilBERT and XLNet; PyTorch for RoBERTa and BigBird.
- Loss Functions: SparseCategoricalCrossentropy for Ten- sorFlow models, CrossEntropyLoss for PyTorch models.
- Optimizers: Adam with learning rates ranging from 2e-5 to 5e-5.
- Epochs: 6 for BigBird (to prevent overfitting on long sequences), 20 for others based on convergence.

H. Evaluation Metrics

All models were evaluated on the same held-out test set using a consistent set of evaluation metrics. These included:

- Accuracy: Overall classification correctness.

- Precision, Recall, F1-Score: Computed per class and averaged using macro and weighted strategies.
- Confusion Matrix: Generated for visualizing misclassification patterns and understanding inter-class confusion.

Metrics were calculated using Scikit-learn functions post-inference, with the predicted logits converted to class indices using argmax.

I. Tooling and Environment

The entire research was implemented using the following toolkits and environments:

- Hugging Face Transformers: Pretrained models and tokenizers.
- Hugging Face Datasets: Dataset loading, efficient tokenization and data processing, and metric computation.
- TensorFlow & PyTorch: Training and Inference of the Model.
- Scikit-learn: Label encoding, metric computation, and confusion matrices.
- Matplotlib & Seaborn: Loss, accuracy and distribution visualizations

All experiments were conducted with the use of GPU due to the quickness of training and reproduction [34].

J. Summary of Methodological Rationale

The methodological setup was developed to make a comprehensive and controlled comparison between transformer models with customized architectures of model. We standardized preprocessing, tokenization, training conditions and evaluation metrics, to the end of ensuring that performance gaps are genuinely clear and obvious by model architecture rather than experimental variability. Our results provide ac- tionable suggestions for researchers and developers in affective computing, particularly those building emotionally intelligent systems in the form of chatbots, mental health assistants and conversational agents.

RESULTS

This section provides an in-depth analysis of performance results from the comparison of four Transformerbased models—DistilBERT, XLNet, RoBERTa, and BigBird—on emotion detection task leveraging conversational text. Models were developed with training and testing performed on identical datasets, with model performance evaluated with a set of standard metrics such as accuracy, precision, recall, F1-score and confusion matrix analysis. Training behaviour and generalisation to unseen data was assessed through performance plots and classification reports.

A. Performance Comparison

1) *Model Accuracy and Loss:* The performance of each model was tracked across training epochs using both accuracy and loss curves. Figures 8,11,9,10 illustrate the training and validation accuracy and loss for each model.

- *DistilBERT* demonstrated rapid convergence, achieving a peak training accuracy of approximately 99.41%, with a corresponding validation accuracy stabilizing around 93.55%. The training loss steadily decreased, while the validation loss exhibited moderate fluctuation, indicat- ing slight overfitting tendencies over extended training epochs. Despite its compact architecture, DistilBERT showed strong classification capability across most emo- tion classes, though performance on less represented categories such as surprise was comparatively weaker (Figure: 8).
- *XLNet* achieved a training accuracy of 99.32% and a validation accuracy reaching 93.90%. Its learning curve remained consistent with a low training loss and a vali- dation loss that gradually increased in later epochs, sug- gesting some generalization limits (Figure 9). The model handled syntactic complexity well, supporting accurate learning of emotion representations embedded in more grammatically rich sentences.
- *RoBERTa* outperformed the other models in terms of validation accuracy, peaking at 94.20%. Both training and validation accuracy remained high and stable throughout 20 epochs of training. The loss graph indicated minimal overfitting, with validation loss staying consistently low and close to the training loss (Figure 10). These results affirm the efficacy of RoBERTa's pretraining strategies and robust context modeling in emotion classification tasks.
- *BigBird*, designed for handling longer sequences, achieved a training accuracy of 94.20% and a validation accuracy of 93.0% over 6 epochs. While the training loss showed a consistent downward trend, the validation loss began to increase slightly in later epochs, reflecting early signs of overfitting on longer text sequences (Figure: 11). Nonetheless, BigBird maintained stable performance in processing extended conversations.

2) *Confusion Matrices analysis:* Confusion matrices were used to examine how well each model performed across

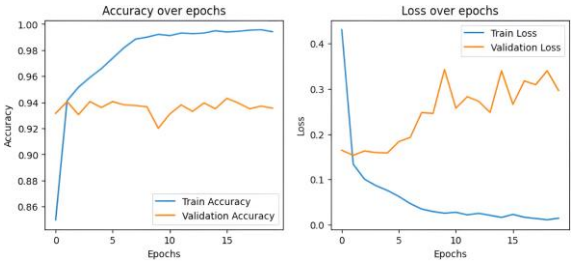

Fig. 8. DistilBERT Training & Validation Accuracy and Loss

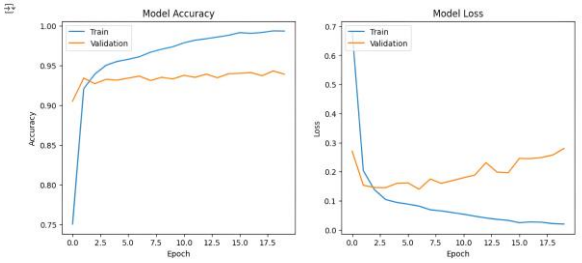

Fig. 9. XLnet Training & Validation Accuracy and Loss

The six emotion categories. Figures:12,13,14,15 display the distribution of true and predicted labels for each model. A comparative observation across models revealed that certain emotion pairs, such as joy and love, or fear and surprise, were more frequently misclassified. This can be attributed to overlapping lexical patterns or shared contextual cues in the input data.

• *DistilBERT* exhibited relatively high accuracy across joy, sadness, and anger, while showing more frequent misclassifications in love and surprise (Figure: 12).

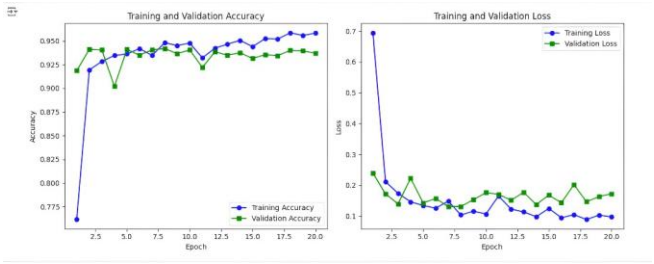

Fig. 10. RoBERTa Training & Validation Accuracy and Loss

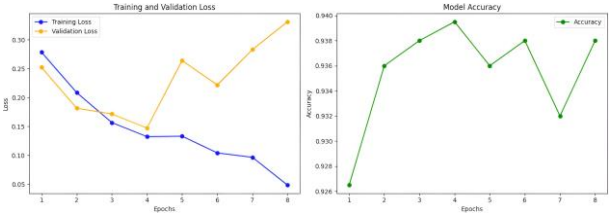

Fig. 11. BigBird Training & Validation Accuracy and Loss

• *XLNet* handled fear particularly well, showing improved recall in this class. However, it also showed occasional confusion between joy and love, as seen in its confusion matrix (Figure 13).

• *RoBERTa* displayed balanced classification across all emotion categories, with fewer misclassifications, espe- cially in closely related emotions such as joy and love. Its confusion matrix showed the most compact diagonal (Figure: 14).

• *BigBird* also delivered balanced predictions, but confu- sion between joy and love was again evident. It managed to preserve context for longer sentences, helping retain high classification integrity across sadness and fear (Fig- ure: 15).

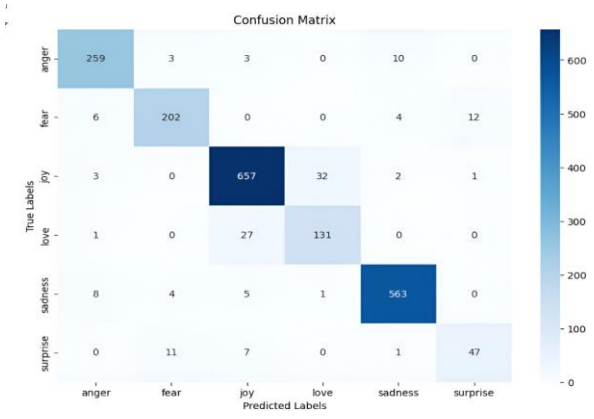

Fig. 12. Confusion Matrix of DistilBERT

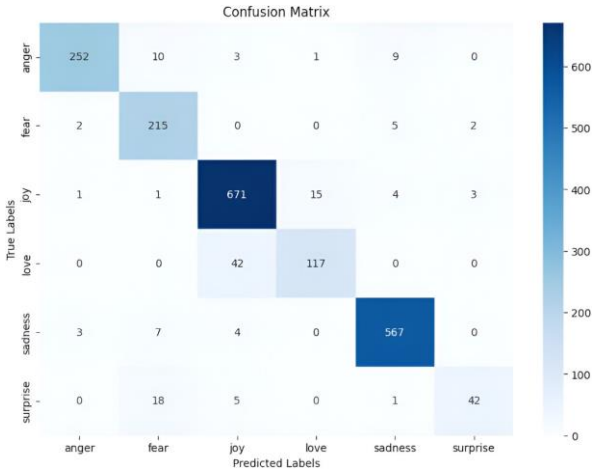

Fig. 13. Confusion Matrix of XLnet

B. Detailed Model Behavior Analysis

1) Interpretation of Prediction Results: To further under- stand model behavior, classification reports were generated for each model, summarizing precision, recall, and F1-score

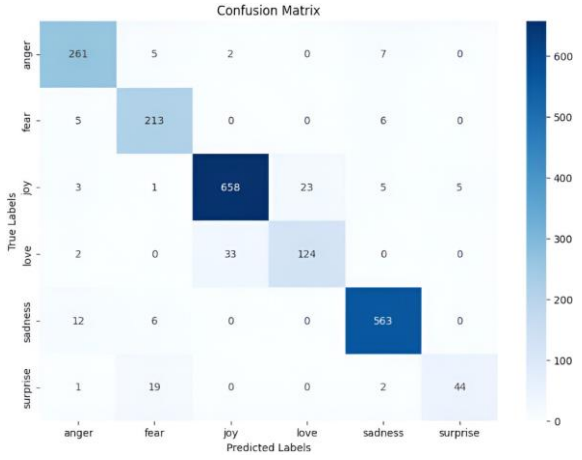

Fig. 14. Confusion Matrix of RoBerta

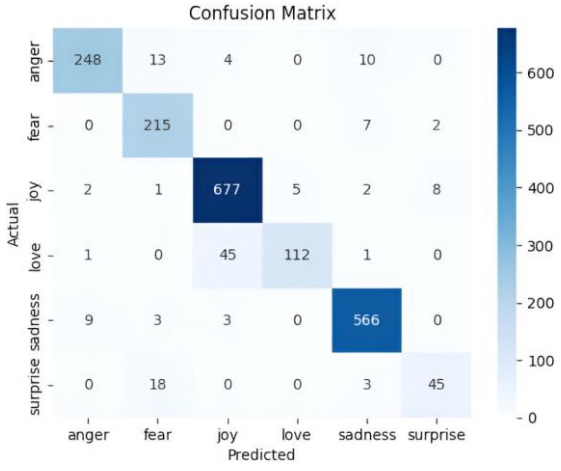

Fig. 15. Confusion Matrix of BigBird

for each class. Figures: 16,17,18,19 provide a visualization of these metrics.

- *DistilBERT* offered consistently high precision and recall for joy, sadness, and anger. However, its classification of surprise suffered from lower recall (0.71), which may be attributed to the low frequency of this label in the training set (Figure: 16).
- *XLNet* demonstrated strong recall for fear (0.96) and sadness (0.98), leveraging its permutation-based attention to learn complex emotional structures in text (Figure: 17). It showed high F1-scores across all major classes, maintaining balanced performance.
- *RoBERTa* achieved the most uniform precision-recall balance across all six emotion categories (Figure: 18). It maintained macro and weighted F1-scores above 0.93, indicating strong robustness and generalization across high- and low-resource classes alike.
- *BigBird* while optimized for long sequences, handled shorter texts effectively as well. Its F1-score for love and surprise remained consistent with other models, indicat- ing its capability to process nuanced emotional signals in longer conversations (Figure: 19).

Classification Report:

	precision	recall	f1-score	support
anger	0.94	0.94	0.94	275
fear	0.92	0.90	0.91	224
joy	0.94	0.95	0.94	695
love	0.80	0.82	0.81	159
sadness	0.97	0.97	0.97	581
surprise	0.78	0.71	0.75	66
accuracy			0.93	2000
macro avg	0.89	0.88	0.89	2000
weighted avg	0.93	0.93	0.93	2000

Accuracy: 0.9295
Precision: 0.929359910867716
Recall: 0.9295
F1 Score: 0.9293456749446702

Fig. 16. Classification of DistilBERT

Accuracy: 0.932

	precision	recall	f1-score	support
anger	0.98	0.92	0.95	275
fear	0.86	0.96	0.91	224
joy	0.93	0.97	0.95	695
love	0.88	0.74	0.80	159
sadness	0.97	0.98	0.97	581
surprise	0.89	0.64	0.74	66
accuracy			0.93	2000
macro avg	0.92	0.86	0.89	2000
weighted avg	0.93	0.93	0.93	2000

Precision: 0.9323623318889548
Recall: 0.932
F-Score: 0.9303454318458936

Fig. 17. Classification of XLnet

2) Strengths and Trade-offs: Each model exhibited strengths based on its architecture and training philosophy. These strengths and potential trade-offs are summarized below:

- *DistilBERT* offered fast training and low computational overhead, making it well-suited for real-time or low- resource applications. However, its reduced model size limited its ability to fully capture subtleties in low- frequency emotion classes.
- *XLNet* excelled at learning deep contextual relationships, benefiting from its permutation-based learning strategy. It required more training time and memory, which may impact deployment in constrained environments.
- *RoBERTa* consistently performed well across all evaluation metrics. Its deep pretraining and optimization strategies contributed to high accuracy but required longer training times and substantial compute resources.
- *BigBird* has ability to work with long inputs and long conversations. Although it was suitable for multiturn

Classification Report:

	precision	recall	f1-score	support
anger	0.92	0.95	0.93	275
fear	0.87	0.95	0.91	224
joy	0.95	0.95	0.95	695
love	0.84	0.78	0.81	159
sadness	0.97	0.97	0.97	581
surprise	0.90	0.67	0.77	66
accuracy			0.93	2000
macro avg	0.91	0.88	0.89	2000
weighted avg	0.93	0.93	0.93	2000

Accuracy: 0.9315
Precision: 0.9313126062408366
Recall: 0.9315
F1 Score: 0.9305215774901531

Fig. 18. Classification of RoBERta

	precision	recall	f1-score	support
anger	0.95	0.90	0.93	275
fear	0.86	0.96	0.91	224
joy	0.93	0.97	0.95	695
love	0.96	0.70	0.81	159
sadness	0.96	0.97	0.97	581
surprise	0.82	0.68	0.74	66
accuracy			0.93	2000
macro avg	0.91	0.87	0.88	2000
weighted avg	0.93	0.93	0.93	2000

Fig. 19. Confusion Matrix of BigBird

dialogues and threaded conversations, it was too heavy on the extraction of subtle emotions expressed over short utterances.

3) *Summary of Performance Metrics:* Table II presents a summary of important metrics for each model: the global accuracy, the macro and weighted F1-scores and some class-wise performance highlights.

TABLE II
KEY PERFORMANCE METRICS FOR DISTILBERT, XLNET, ROBERTA, AND BIGBIRD

Model	Accuracy	Macro F1	Weighted F1	Notable Class Strengths
DistilBERT	92.95%	0.89	0.93	Joy, Sadness
XLNet	93.20%	0.89	0.93	Fear, Joy
RoBERTa	93.15%	0.89	0.93	BBalance across all
BigBird	93.00%	0.88	0.93	Sadness, Long Input (required)

DISCUSSION

The comparative analysis of DistilBERT, XLNet, RoBERTa, and BigBird models for detecting emotions in conversational text demonstrates both the strengths and weaknesses of contemporary Transformer-based models for handling affective information. Such models, with varying design philosophies and optimization schemes, provide insights into how deep language models understand and categorize human emotions conveyed in context through natural language interaction. The results of this study validate the applicability of Transformer-based models for emotion classification tasks, but point out the crucial issues for further development and investigation.

The implementation of DistilBERT, XLNet, RoBERTa, and BigBird has provided great knowledge about how Transformer models are-stored-how better and more efficient they become in perceiving and reasoning over emotional content during text conversations. This section examines the details of the explained results including the captured emotional dynamics, the difficulties faced and the approach taken by the different models in the detection of emotions.

CONCLUSION

The goal of this research was to assess the suitability and performance levels of four different Transformer models: DistilBERT, XLNet, RoBERTa and BigBird as applied to emotion detection in conversational texts. The detailed analysis revealed some big insights, with every model having its own specific strengths and weaknesses in the emotion detection space. This study revealed that RoBERTa clearly outperformed other models in overall accuracy, performing especially well with subtle emotions like sadness or joy due to its richer context knowledge. The BigBird was very good at handling longer passages of text, thus preserving the context for longer dialogues like in customer service or therapy conversations. DistilBERT, while less robust in depth, offered significant efficiency, making it ideal for real-time applications. XLNet's advanced handling of complex sentence structures enabled superior performance in detecting emotions embedded within intricate expressions. However, all models shared a common challenge in accurately identifying the emotion 'surprise,' highlighting a potential area for model refinement and further research.

Contributions to the Field

This research significantly advances the understanding of emotion detection within AI and NLP fields by:

1. Demonstrating the feasibility and effectiveness of using advanced Transformer models for emotion detection in varied textual conversations.
2. Highlighting the specific strengths and weaknesses of each model, providing a roadmap for future applications and improvements.
3. Identifying key challenges in emotion detection, particularly in distinguishing between closely related emotional states, which can inform subsequent model training and algorithm adjustments.

The findings of this study not only enhance existing methodologies in affective computing but also contribute to the broader discourse on improving human-computer interaction. By refining AI's ability to understand human emotions accurately, this research supports the development of more empathetic and responsive AI systems, paving the way for innovations that could revolutionize customer service, mental health therapies, and social media analytics.

Future Work

The potential for advancing the field of emotion detection through Transformer models is vast, with numerous av- enues for further development and application. This research has laid a robust foundation, demonstrating the efficacy of DistilBERT, XLNet, RoBERTa, and BigBird in understanding emotional dynamics within textual data. Building upon these findings, future work can focus on refining these models, exploring extended applications, and implement- ing real-time systems.

REFERENCES

1. [1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," 2019. [Online]. Available: https://arxiv.org/abs/1905.02947

[2] P. Pereira, H. Moniz, and J. P. Carvalho, "Deep emotion recognition in textual conversations: a survey," Artificial Intelligence Review, vol. 58, no. 1, p. 10, Nov 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-11010-y

[3] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "Roberta- lstm: A hybrid model for sentiment analysis with transformer and recurrent neural network," IEEE Access, vol. 10, pp. 21 517–21 525, 2022.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, XLNet: generalized autoregressive pretraining for language understand-ing. Red Hook, NY, USA: Curran Associates Inc., 2019.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[7] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang et al., "Big bird: Transformers for longer sequences," Advances in neural information processing systems, vol. 33, pp. 17 283–17 297, 2020.

[8] D. Cortiz, "Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta,

xlnet and electra," 2021. [Online]. Available: https://arxiv.org/abs/2104.02041

[9] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, "Neural architecture search for transformers: A survey," IEEE Access, vol. 10, pp. 108 374–108 412, 2022.

[10] R. Thapa, B. Lamichhane, D. Ma, and X. Jiao, "Spamhd: Memory- efficient text spam detection using brain-inspired hyperdimensional computing," in 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2021, pp. 84–89.

[11] T. Kaneva, B. Evstatiev, I. Valova, N. Valov, and K. Gabrovska- Evstatieva, "Comparing different evaluation metrics with the grid search method for classification of highly imbalanced data," in 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2024, pp. 1–5.

[12] T. Kim and P. Vossen, "Emoberta: Speaker-aware emotion recognition in conversation with roberta," 2021. [Online]. Available: https://arxiv.org/abs/2108.12009

[13] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," Trans. Multi., vol. 26, p. 776–788, Jan. 2024. [Online]. Available: https://doi.org/10.1109/TMM.2023.3271019

[14] D. Cortiz, "Exploring transformers models for emotion recognition: a comparision of bert, distilbert, roberta, xlnet and electra," in Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System, ser. CCRIS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 230–234. [Online].

Available:https://doi.org/10.1145/3562007.3562051

[15] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter, Probabilistic Networks and Expert Systems, 01 2001, vol. 43.

[16] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," Information Fusion, vol. 59, pp. 103–126, 2020.

[17] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," Journal of Intelligent Systems, vol. 31, no. 1,pp. 113–126, 2022. [Online]. Available: https://doi.org/10.1515/jisys-2022-0001

[18] H. C. W. Acheampong, Francisca Adom Nunoo-Mensah, "Transformer models for text-based emotion detection: a review of bert-based approaches," Artificial Intelligence Review, vol. 54, no. 8, pp. 5789–5829, Dec 2021.

[19] N. Mughal, G. Mujtaba, S. Shaikh, A. Kumar, and S. M. Daudpota, "Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis," IEEE Access, vol. 12, pp. 60 943–60 959, 2024.

[20] M. V. Koroteev, "Bert: A review of applications in natural language processing and understanding," 2021. [Online]. Available: https://arxiv.org/abs/2103.11943

[21] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021. [Online]. Available: https://arxiv.org/abs/2106.04554

[22] B. Mahesh, "Machine learning algorithms -a review," International Journal of Science and Research (IJSR), vol. 9, 01 2019.

[23] S. Salam and R. Gupta, "Emotion detection and recognition from text using machine learning," International Journal of Computer Sciences and Engineering, vol. 6, pp. 341–345, 06 2018.

[24] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," Neural Computation, vol. 31, no. 7, pp. 1235–1270, 2019.

[25] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud, "Text-based emotion recognition using deep learning approach," Computational Intelligence and Neuroscience, vol. 2022, no. 1, p. 2645381, 2022.

[26] D. Haryadi and G. P. Kusuma, "Emotion detection in text using nested long short-term memory," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100645

[27] C. Zhang, "Natural language processing: Classification of web texts combined with deep learning," Journal of ICT Standardization, vol. 13, no. 1, pp. 25–40, 2025.

[28] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[29] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in 2020 17th International Computer Conference on Wavelet

Active Media Technology and Information Processing (ICCWAMTIP), 2020, pp. 117– 121.

[30] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[31] J. Xiao, Z. Shao, S. Han, and Z. Li, “A study on sentiment analysis of online course review information based on xlnet-bigru,” in 2022 12th International Conference on Information Technology in Medicine and Education (ITME), 2022.

[32] P. Kameswari and P. Sudha Rani, “Multilingual spam classification using advanced deep learning techniques,” in 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), 2024, pp. 1595–1601.

[33] C. Dhivyaa, K. Nithya, G. Sendooran, R. Sudhakar, K. Kumar, and S. Kumar, “Xlnet transfer learning model for sentimental analysis,” in 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), 2023, pp. 76–84.

[34] D. Rothman, Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more, 2021.