Emotion Detection In Text Using Natural Language Processing

Diaa Salama AbdElminaam[1], Noor Hatem Hassan[2], Nourhan Amr Mohamed[3],
Shaden Essam[4], Abdullah Mohamed[5],
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], noor2206987[2],
nouran2205080[3], shaden2208768[4], abdullah2207019[5]{@miuegypt.edu.eg}

***Abstract*—Emotion detection is an essential task in natural language processing (NLP), aiming to identify the emotion from the text.However, Emotion detection remains a complex problem due to the nature of emotional expression and because researchers focus on other detections rather than emotion detection from text.This research aims to develop a system using machine learning techniques and deep learning techniques. The main objective is to compare the performance of different models to classify emotion categories in text given. We used 7 machine learning algorithms MultinomialNB , Logestic Regression, Gradient Boosting, Linear SVM, K-Nearest Neighbors, Decision Tree, MLP classifier and 7 deep learning models LSTM, BILSTM, GRU, BiGRU, 1D_CNN, CNN_BILSTM, CNN_Attention applied to two different emotion detection datasets including text features and emotion label for both dataset preprocessing where applied involving checking for missing values , removing all no alphabetic characters, conversion of text to lowercase, removing of commen english stopwords and stemming words to their root.Models were evaluated using performance metrics such as accuracy, precision, recall and F1-score. The results showed that models BI-GRU 0.93% and SVM 0.88 % have the highest accuracy in one of the datasets and the other dataset showed that Linear SVM 0.87% and 1D_CNN 0.93%. The study concludes that using models with highest performance improves emotion detection performance.Future work will focus on hyperparameter improvement to optimize model effectiveness.**

Keywords: Emotion Detection, Machine Learning, Deep Learning, MultinomialNB , Logistic Regression, Gradient Boosting, Linear SVM, K-Nearest Neighbors, Decision Tree, MLP classifier , LSTM, BILSTM, GRU, BiGRU, 1D_CNN, CNN_BILSTM, CNN_Attention

I. INTRODUCTION

In the age of digital technology, our online world relies on understanding how people feel through their words. Whether you're offering mental health support, running customer service, analyzing social media chatter, or designing human–computer interactions, picking up on emotional tone in text opens up countless possibilities. Traditional approaches to emotion recognition depend on physical cues such as facial expressions or vocal intonation, but these are absent in text-based communication. As more conversations shift online, we need new methods to bridge this gap. Natural Language Processing (NLP), together with Machine Learning (ML) and Deep Learning (DL), provides a powerful pathway to extracting emotional meaning from text with high accuracy.

It can be difficult to identify emotions in plain text. Words frequently reflect individual expression, have multiple meanings, and are strongly influenced by context. Emotion detection seeks to identify particular emotions like joy, anger, sadness, fear, surprise, or disgust, in contrast to sentiment analysis, which simply categorizes text as positive or negative. Although both deep learning (DL) and machine learning (ML) models have been used to address this issue, few studies have directly compared them in the same place. To ensure a fair and meaningful comparison of their performance, even fewer have employed a standardized set of evaluation metrics, including accuracy, precision, recall, and F1-score.

The main goal of this project is to evaluate and compare the effectiveness of modern deep learning models and traditional machine learning algorithms in the task of textual emotion classification. In terms of machine learning, we look at popular models such as the Support Vector Machine (SVM), Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Multilayer Perceptron (MLP). Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU), 1D Convolutional Neural Network (CNN), and hybrid models like CNN-BiLSTM ,CNN-LSTM and CNN with attention mechanisms were the sequential models that we concentrated on for the DL approach. To guarantee a fair and useful comparison, all models were trained using the same preprocessing methods and evaluated using the same performance metrics.

From the perspective of machine learning, these algorithms frequently depend on unique textual features. Raw text is converted into numerical vectors using methods like TF-IDF or Bag-of-Words, which are later fed into classifiers like SVM or KNN. These models are fast to compute, accessible, and reasonably light. The deeper contextual and sequential relationships needed for understanding emotions in language are often difficult for them to capture.

On the other hand, deep learning models are made to automatically extract layered, abstract features from unprocessed input data. Sequential dependencies within sentences are best handled by constant neural networks like LSTM and GRU, while bidirectional neural networks enable models to simultaneously learn context from past and future tokens. Attention mechanisms further improve performance by enabling models to concentrate on emotionally important sections of the text, while CNN layers helps in capturing local patterns such as emotional phrases. Regardless of having a higher computational cost, these benefits make DL models ideal for emotion detection.

This study's main research question was whether deep learning models could accurately detect emotional details in raw text than traditional machine learning methods. To address this, we used a sizable labeled dataset with clear preprocessing methods, such as one-hot encoding of emotion labels, tokenization, padding, stopword removal, and rare word filtering. We then trained and tested each model under the same circumstances to observe the impact of its architecture and learning style on performance.

As a summary of our contributions:

- For multi-class emotion classification, we thoroughly compared traditional machine learning and advanced deep learning models.
- To guarantee high-quality input for every model, we used reliable and efficient preprocessing techniques.
- We highlighted the advantages of each method and showed how different models react to text's emotional nuances.
- We showed that some deep learning models perform noticeably better than others at capturing contextual emotion, particularly those with attention mechanisms and bidirectional layers.
- In order to simplify future research and practical application, we offer a consistent framework for evaluating emotion detection models.

The rest of this paper is structured as follows: Section II discusses related work in the field of emotion recognition. Section III explains our methodology, including dataset details, preprocessing techniques, and model architectures. Section IV presents and analyzes our experimental results. Finally, Section V concludes the study and suggests directions for future research.

II. Related Work

These researches on Emotion Detection in text using NLP is based on the work experienced researchers. These insightful studies have helped us to understand the problem and using the right technique to improve it. These influential works are discussed in the following section. Each will be properly acknowledged and cited in our references.

This research paper focuses on text-based emotion detection from Reddit comments. [1],. They use GoEmotions dataset containing 27 emotions mapped into the basic 6 (joy,sadness,anger,fear,disgust,surprised). To pre-process they use with the basic techniques and they use lemmatization to shorten words. They used 6 different machine learning models, 3 Ensemble and 1 deep learning, they stated that the Stacking Classifier: an advanced ensemble method,combines multiple base models and a meta-classifier was the model with highest performance overall.The best-performing was deployed in a Streamlit web application to enable real-time, user-facing emotion detection.

A research paper examines how machine, Deep and ensemble learning algorithms may be used to detect emotions in text [2]. They collect data for the dataset from Twitter posts. They classified emotions into four categories: joy, anger, fear, and sadness. In terms of preprocessing. they applied Lemmatization and stemming Removal of punctuation, HTML tags, URLs, and stop words and Expansion of abbreviations (e.g., "omg" to "oh my god"). They used Bag-of-Words and Word Embedding and compared the results for feature extraction but the decided to use word embedding because it preformed better with deep learning models. They used this machine learning model: Logistic Regression, SVM, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest ,Deep Learning models: CNN , LSTM, GRU, BiLSTM, BiGRU and A final ensemble model combining BiLSTM and BiGRU. In conclusion, BiGRU and BiLSTM ensembles have the best accuracy

This research paper focuses on text-based emotion detection using Natural language processing and Neural Networks.[3] .They used custom dataset containing whatsapp status text categorized into 6 emotions: joy, sadness,anger, fear, disgust and surprise.For preprocessing they did stop-word removal, eliminate punctuation , emojis and special characters and they tokenized the text.They used 3 different approaches: NRCLex Lexicon based , NLP based and deep learning neural network method.Firstly, the NRCLex method used the NRC lexicon to check whether the word exist or not and to map words to 8 basic emotions and sentiments.Secondly, The neural network model uses RNNs (Recurrent Neural Network) they are designed for text data and used GRU to focus on classifying text into 8 basic emotions.Moreover, the NLP method using traditional text processing and classification techniques.They evaluated this three models and the neural network model achieved the highest accuracy of 99.56 , compared to the NLP based model (83.36) and the NRCLex method (64.44).

The Objective of this paper is to present a comparative study on deep learning models for emotion detection from text. [4] They compare the preformances of Recurrent Neural Networks (RNNs) — LSTM, BiLSTM, and GRU on their dataset(ISEAR (International Survey on Emotion Antecedents and Reactions)) which is classified to 7 emotion classes: anger, sadness, disgust, shame, fear, joy, and guilt. In Terms of Preprocessing Feature Engineering, they used Tokenization and lemmatization, Word embeddings were generated using a pre-trained Word2Vec model an they removed stop-words and punctuation.Regarding their performance, GRU was the best-model with accuracy 60. In conclusion, although GRU accuracy was not that high due to limited dataset but still a good performance model to be used in detection

This reasarch paper focuses on text based emotion detection using Natural language processing and deep learning,[5] The database the author creted it from Friends Tv show with each utterance labeled with 7 emotions: joyful, sad,mad, scared, powerful, peaceful, and neutral.They used four sequence based CNN models (SCNNc,SCNNv,SCNNa_c,SCNNa_v) among them all SCNNa_c acheved the highest performance with (37.9%) for 7 emotion classes and (54%) for 3 emotion categories.

This paper docus in detecting emotion in social media text by combining NLP and machine learning[6]The author collected tweets labeled with 6 emotions.They cleaned the data by removing unwanted part and using lemmatization. There approch was firstly using NLP alone then using machine learing method(SMO and J48) alone then combining them together. After combination they tested their model on 900 tweets and achecved 91.7% using SMO and 85.4% using J48.
This research [7] highlights how important it is to detect emotions from text, especially forimproving interactions between humans and machines. The authors address the challenge of understanding emotions like joy, sadness, anger, and love from written language, which is often complex and subtle. Their goal was to compare how well traditional machine learning methods perform against modern deep learning techniques. The study found that deep learning models, especially Bi-GRU, delivered the best results with an accuracy of 78.70%.The discussion shows that these models are more effective at capturing the flow of language. In conclusion, deep learning is a better fit for emotion detection in text, and future work could further enhance results with more advanced models and bigger datasets.

This research [8] highlights how important it is for social robots to understand human emotions in order to interact more naturally and effectively. The main problem the authors tackle is detecting emotions from text in real-time, which is challenging but crucial for meaningful communication. Their goal was to design a framework that uses powerful language models (transformers) along with an emotion-based ontology (EMONTO) to help robots recognize and store emotional responses. In their tests, the system performed well, nearly matching a top-performing model from NVIDIA. The discussion points out ways to improve, like handling imbalanced data and adding input from other sources like voice or images. Overall, the study shows that combining emotion detection with smart data organization can make robots more responsive and emotionally aware.

This paper[9] introduces EmoTxt, a handy open-source tool that can detect emotions like joy, anger, love, and sadness in written text—something most sentiment analysis tools can't do. The authors noticed a gap: while many tools can tell if a message is positive or negative, very few can actually understand the emotion behind it. Their goal was to create a flexible toolkit that not only detects emotions but also lets users train it with their own data. They tested EmoTxt on real conversations from platforms like Stack Overflow and Jira, and it performed really well, with some emotions hitting accuracy scores over 90

This paper [10] presents DLSTA, a deep learning model that can recognize human emotions from written text alone. The model learns to recognize emotional cues like joy, anger, or sadness that are hidden in everyday language by combining word embeddings with natural language processing techniques. In order to gain a deeper understanding of context and emotions, it includes information from questionnaires in addition to words. When compared to other well-known techniques, DLSTA performed very well, identifying emotions with an accuracy of 98 percent. Whether in social media, mental health, or education, this work holds great promise for creating more emotionally intelligent systems.

The research presents a model called DeepEmotex [11] that detects emotions in text using powerful pre-trained language models like BERT and USE. The authors collected over 500000 emotion-labeled tweets and used them to fine-tune these models, helping them understand and classify emotions like joy, anger, and sadness more accurately. DeepEmotex BERT achieved high accuracy with 92 percent on test data and 73 percent on benchmark datasets, significantly outperforming older models like BiLSTM. The study shows that using more labeled data for fine-tuning improves performance and reduces errors, making DeepEmotex a strong method for emotion detection in text.

The study [12] examines how well different machine learning models can detect emotions from text. It focuses on models like Logistic Regression, Extra Randomized Trees (ERT), Voting Classifier, SGD, and LinearSVC, and tests them using feature extraction methods such as TF-IDF, Bag-of-Words, and N-grams on two datasets: ISEAR and AIT-2018. The results show that using N-grams with LinearSVC and ERT gives the best accuracy (88.63 and 89.14, respectively), showing how important word sequence patterns are for recognizing emotions. The study also finds that single models often perform better than ensemble methods like the Voting Classifier in this task. Overall, it provides useful insights into choosing the right models and features for emotion detection in text.

III. Proposed Methodology

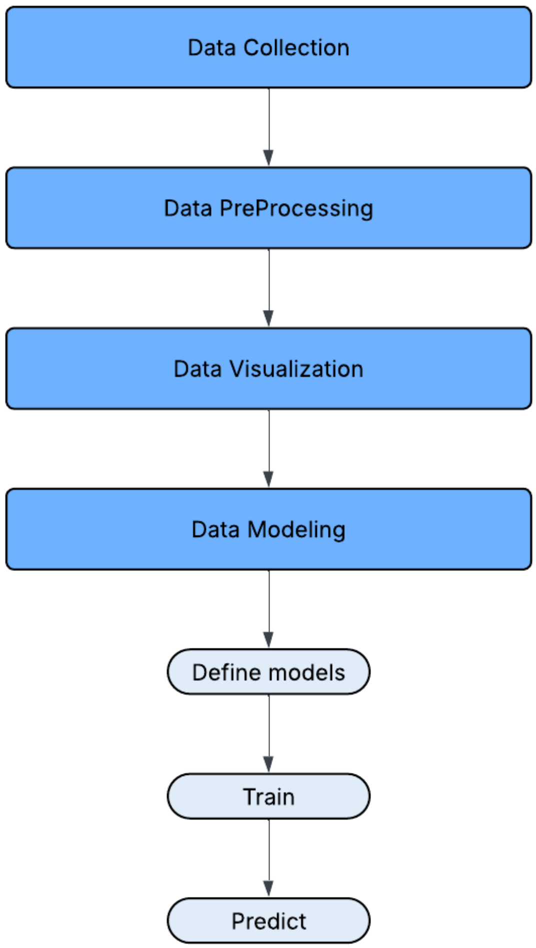

Fig. 1. Your caption here

Fig. 2. Steps

A. Datasets Descriptions

The first dataset [13] used in this study is called the Emotions Dataset for NLP + Neutral Emotion, created by Thuan Naheem Pakeer and published on January 1, 2024 .it includes Approximately 21,000–22,000 labeled text samples,each labeled with one of this 7 Emotions:'sad' 'anger' 'love' 'surprise' 'fear' 'joy' 'neutral'. This data set is useful for the detection of emotions and the addition of neutrals allows one to detect the neutral emotion in text, which is overlooked a lot because it is very important to detect neutral emotions for realistic emotion detection.using trained models to calculate the their accuracy and predict emotions in a given text.

TABLE I
Features of dataset 1

Features	Type	Description
ID	Numerical	ID
Subtitle	Categorical	Text input
Emotion	Categorical	Emotion category label

The second and final dataset [14] Is Sentiment and Emotion Analysis Dataset. Created by Kaggle User Kushagara and published on December 11, 2024 focuses on both sentiment and emotion classification but in our project we used emotion classification. This dataset uses over 422,000 sentences labeled with six distinct emotions as follows: Joy (143,067), Sadness (121,187), Anger (59,317), Fear (49,649), Love(34,554), Surprise(14,972). This data is perfectly suited for tasks like emotion detection in text , multitask nlp applications and Pre-training or fine-tuning transformer models like BERT, GPT, or similar architectures. Each sample consists of a sentence and its corresponding label ensuring ease of use.

TABLE II
Features of dataset 2

Features	Type	Description
Sentence	Categorical	Text input
Emotion	Categorical	Emotion category label

B. Data Preprocessing

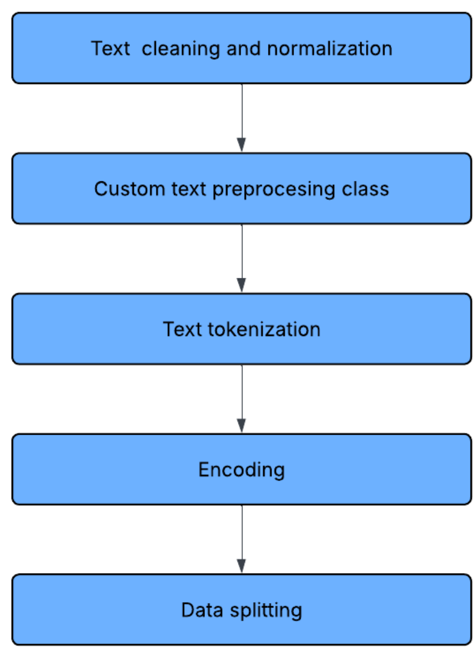

Fig. 3. Preprocessing Steps

One of the core concepts of data science is data preparation, which is necessary to get datasets ready for machine learning and deep learning models to use and evaluate. During this crucial stage, a number of actions are taken to improve the quality of the data and make sure it is appropriate for analysis.

1) **Data Loading:**
The first step involved loading the emotion labeled text dataset which contains sentences labeled with different emotion classes.

2) **Text Cleaning and normalization:**
To ensure accuracy and consistency in our analysis, the multiple normalization steps were applied: Removal of all non alphabetic characters, conversion of text to lowercase, Removal of commen english stopwords and stemming of words to their root. These step is done to make sure that the text is clean, consistent and simplified for better model performance.

3) **Custom Text Preprocessing class:**
A reusable preprocessing class was created to use the cleaning logic. This made it easy to apply the same preprocessing across the 2 datasets.

4) **Text Tokenization:**
Before passingthe data to deep learning models, we convert them into tokens using a tokenizer. each word is assigned a unique number Any unknown word is changed with out-of-vocabulary token “<OOV>” . Then we applied post-padding to make sure all are the same length because deep learning models require the same length inputs. This padding technique adds zeros at the end of shorter sequences. The final output is a numerical representation of the text input.

5) **Encoding:** to prepare data for modeling,we used label encoding, which converts each categorical emotion (such as joy, anger or neutral) into a unique integer value (0,1,2,.),because machine models require a numerical input for classification problems.For machine learning models,we applied the TF-IDF (Term Frequency–Inverse Document Frequency) to convert our text inputs into vectors, it captures the important words in text. For Deep learning models:Emotion labels were converted into one-hot encoded vectors, which represent each class as a binary vector
6) **Data Splitting:** The data was spit into training, validation, and testing sets using stratified sampling to preserve the distribution of emotion classes. 80 for training and 20 for validation and testing

C. *Data Visualization*

Data visualization plays crucial role in the data science process by acting as link between intricate datasets and clear conclusions. Visualization assists in identifying patterns trends and outliers. These may not be immediately noticeable in tabular formats. By converting raw data into graphical representations it makes it easier to comprehend the data more deeply. This leads to better decision-making and more successful result communication.

1) **Emotion categories distrubution :**
As shown in Figure 5, we analyze the number of each category in emotion detection dataset.The x-axis corresponds to emotion classes and Y-axis corresponds to count of each emotion.

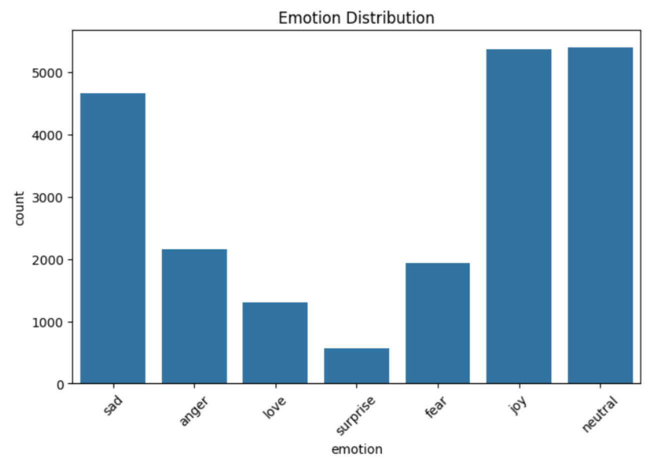

Fig. 4. Emotion distribution in Dataset 1

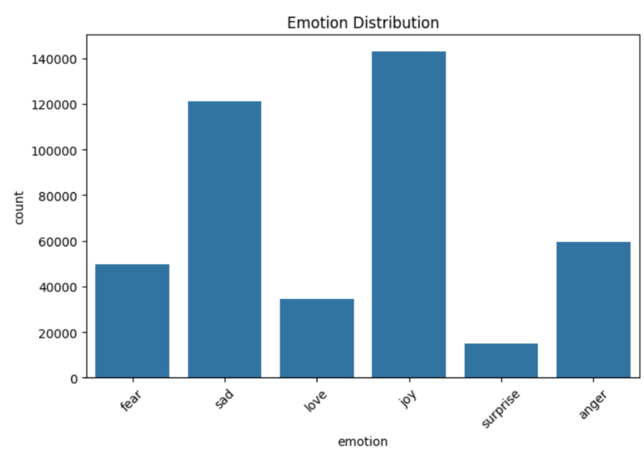

Fig. 5. Emotion distribution in Dataset 2

2) **Text length distribution:**
The graph shows 7 Text length distribution calculates the number of words in each processed text entry. The x-axis represent text length and y-axis represent the count.

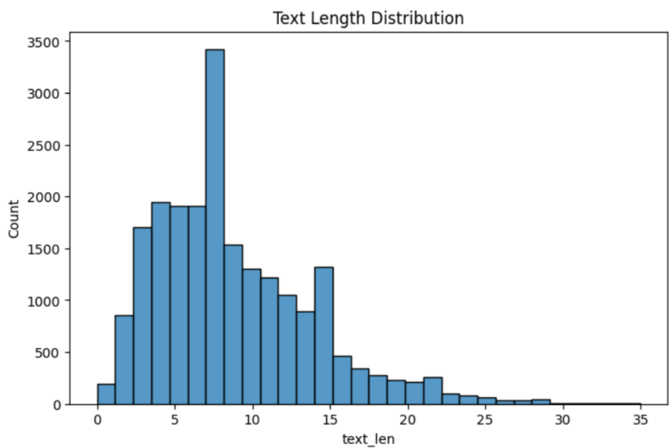

Fig. 6. Text distribution in dataset 1

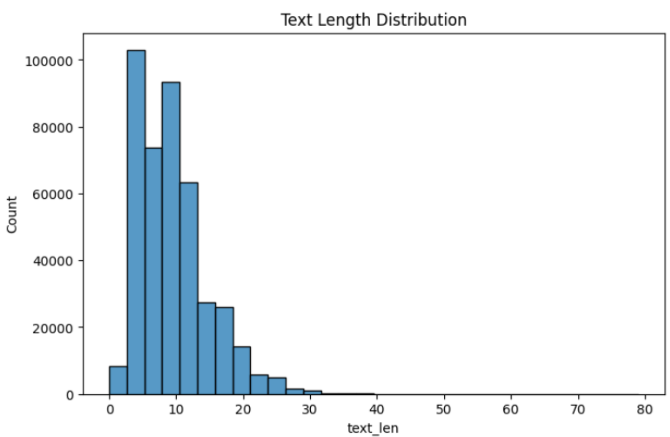

Fig. 7. Text distribution in dataset 2

3) **Word cloud**
The figure 9 generates word cloud from the combined processed text data. All text entries are merged into one string to visualize the most frequent words. It helps quickly identify commen terms in dataset.

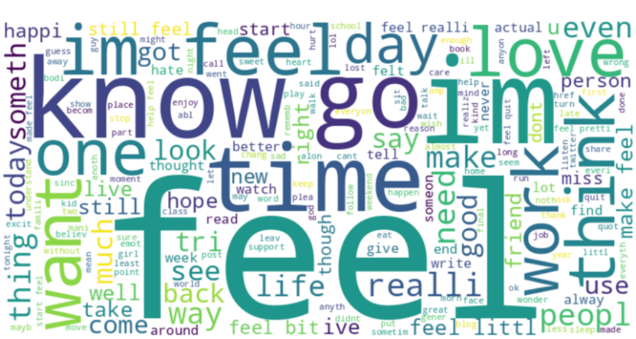

Fig. 8. Word cloud in dataset 1

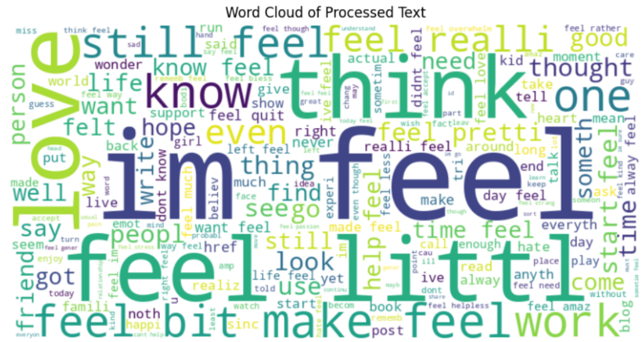

Fig. 9. Word cloud

D. *Used Algorithms*

To classify emotions in an accurate way from text and words, different machine learning and deep learning models were used. Based on how it processes language patterns, context and emotions each and every model contributes differently. The following demonstrate the models, their working , and how they improve the emotion detection.

1) **Logistic Regression :**
Logistic Regression is a classification algorithm that is used widely. It works on the categorical outcome as it models the probability of it. In our project, it was used to predict the emotions as a categories (such as joy, anger, sadness, etc.) from text. Although it has regression in its name but it is not used for regression tasks. It is used for binary or multi-class classification problems.
Logistic regression works by applying the sigmoid function to a linear combination of input features to calculate the probability that a text belongs to a an emotion class. Logistic Regression performed well in this project because it is somehow simple, fast, and has an ability to handle linearly separable data. It helped us understand which features in text contribute most to predicting specific emotions.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \qquad (1)$$

2) **Naive Bayes (MultinomialNB):**
Naive Bayes is based on Bayes' theorem and it is a probabilistic classifier. It says that features are independent. The Multinomial version is particularly effective for text classification tasks where features represent discrete counts, such as word frequencies.

$$P(C_k \mid x) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(x)} \qquad (2)$$

3) **K-Nearest Neighbors (KNN):**
K-Nearest Neighbors is a an algorithm which is non-parametric classification one that predicts the class of a new instance by identifying the majority class among its 'k' closest neighbors in the training data. k is a constant number How it works: For a given sentence, the model calculates its distance (usually Euclidean) from all training sentences, finds the k nearest ones, and predicts the emotion class based on the majority .
Impact on the Project: KNN was helpful for capturing local patterns in text embeddings. Although it's slower with large datasets and sensitive to feature scaling, it allowed us to explore how similar expressions lead to similar emotional interpretations. It provided a useful contrast to more complex models.
4) **Decision Tree:**
Decision Trees is an algorithm that splits the data into branches according to feature values, forming a tree-like structure where each internal node represents a decision rule and each leaf node represents an outcome.
How it works: For example, the model may first check for the presence of a strong negative word, after that split based on other features like intensity or negative words, until reaching a final emotion class.
Impact on the Project: Decision Trees provided easy-to-understand rules for emotion classification. It helped highlight how specific words or combinations lead to certain emotional interpretations, making the results more interpretable.
5) **Linear SVM:**
Linear SVM is a model that tries to separate different types of data by drawing a straight line (or a plane in more dimensions) between them. In our project, it helped us figure out the emotion behind a sentence by learning patterns in the words and how they're used.
After turning each sentence into numbers using techniques like TF-IDF, the model looks for the best way to split the emotions apart—such as anger, joy, or sadness—based on how different they are. What makes Linear SVM strong is that it focuses on the most important examples when deciding where to draw the line, which helps it make more accurate predictions.
It worked really well in our system because it handles text data efficiently and can tell emotions apart even when the differences are subtle. That made it one of the most dependable models we used in detecting emotions from text.
6) **Gradient Boosting:**
Gradient Boosting is another method that builds models sequentially, where each model improves the output accuracy by learning from the previous one and correcting its errors It uses decision trees as base learners and optimizes the model using gradient descent.
Impact on the Project: Gradient Boosting improved the performance of emotion classification by focusing on difficult cases where other models failed. It helped increase accuracy and was especially useful in cases where emotion signals were subtle or mixed within the sentence.
7) **MLP Classifier:**
The MLP Classifier is a neural network which is feedforward type, it includes one or more hidden layers . Each neuron computes a weighted sum of its inputs and passes it through a non-linear activation function. It helped in our project by recognizing emotions in text as it finds deeper patterns because it understands more complex relationships in the data
8) **LSTM (Long Short-Term Memory):**
LSTM is a on of the deep learning models, it is a type of neural network, its good at understanding the order of words in sentence as it remembers each word and connect the words with each other to connect a sentence so in our project it is very useful as it helps in understanding the emotions from the whole sentence
9) **BiLSTM (Bidirectional LSTM):**
BiLSTM does what the LSTM does but its better as it reads the sentence both forward and backward so this gives more better detection of emotions as the understanding of the word usually depends on what comes before it and what comes after it
10) **GRU (Gated Recurrent Unit):**
GRU is like the LSTM but it is faster version it also learns from the sequence of words but with fewer steps so it is faster to train, in our project GRU was more useful in detecting the emotions in the shorter sentences
11) **BiGRU:**
It is a better version of the GRU as it reads the sentence in both directions so it helps in detecting the emotions better like in the BiLSTM because the understanding of a word depends on what comes before and after it
12) **1D CNN (One-Dimensional Convolutional Neural Network):**
1D CNN works on spotting meaningful patterns form phrases or keywords in small groups of words it helped us in our project in detecting a pattern very fast especially in short sentences
13) **CNN-BiLSTM):**
This model combines two strong models and the complete each other as CNN finds the patterns quickly from short couple of words and the BiLSTM complete after it to understand the whole meaning of the big picture
14) **CNN-Attention):**
CNN-Attention is more smarter and has something special as it doesnt look at the whole sentence it learns instead to focus on the most important words and this made our emotion detection more accurate

E. Performance Metrics

We used four performance metrics which are accuracy, precision , recall and F1 score are used to evaluate the model.Accuracy tells whether or not the model prediction is accurate.Recall tells the model how many of the actual positive cases the model was able to detect.Precision tells how many of the predicted positive is correct. F1 score is balance between recall and precision.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

center

IV. Results and Analysis

The following tables show the results for the machine learning and deep learning models used in this study: in ***dataset 1***,In ML models, the highest performance model Linear SVM with accuracy 0.87 and IN DL models, is CNN with accuracy 0.93 in ***dataset 2***,In ML models, the highest performance modelis linear SVM with accuracy 0.878 and IN DL models, is BIGRU with accuracy 0.935 .

TABLE III
Linear SVM Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.84	0.85	0.84	345
Fear	0.84	0.79	0.81	310
Joy	0.85	0.89	0.87	858
Love	0.76	0.62	0.68	209
Neutral	0.94	0.97	0.95	864
Sad	0.89	0.90	0.89	747
Surprise	0.75	0.59	0.66	91
Accuracy	0.8732			
Macro Avg	0.84	0.80	0.82	3424
Weighted Avg	0.87	0.87	0.87	3424

Description: Table III shows the Linear SVM performance on Validation data Set 1. The model achieved an overall accuracy of **87.32%**. Precision, recall, and F1-scores are reported for each emotion class, with the highest performance observed for *Neutral* and *Sad*, and lower scores for *Love* and *Surprise*. **Support Vector Machines (SVM)** were chosen for this task due to their strong generalization ability in high-dimensional spaces, such as text data transformed via TF-IDF(Text Frequency-Inverse Document Frequency).What makes Linear SVM strong is that it focuses on the most important examples when deciding where to draw the line. .

TABLE IV
Naive Bayes Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.80	0.61	0.69	345
Fear	0.83	0.59	0.69	310
Joy	0.65	0.88	0.75	858
Love	0.75	0.35	0.48	209
Neutral	0.86	0.77	0.81	864
Sad	0.71	0.84	0.77	747
Surprise	0.87	0.14	0.25	91
Accuracy	0.7377			
Macro Avg	0.78	0.60	0.63	3424
Weighted Avg	0.76	0.74	0.73	3424

Description: Table IV shows the Naive Bayes classification performance on Validation Set 1. The model achieved an overall accuracy of **73.77%**. Precision, recall, and F1-scores reported for each emotion class, with the highest performance observed for *Joy*, *Neutral*, and *Sad*, while lower F1-scores were obtained for *Love* and *Surprise*. **Naive Bayes** was selected for its simplicity, fast training time, and effectiveness in text classification tasks where feature independence is assumed, such as bag-of-words or TF-IDF representations like we used in this study

TABLE V
Logistic Regression Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.84	0.81	0.82	345
Fear	0.86	0.79	0.82	310
Joy	0.82	0.90	0.86	858
Love	0.72	0.59	0.65	209
Neutral	0.94	0.96	0.95	864
Sad	0.88	0.89	0.89	747
Surprise	0.80	0.53	0.64	91
Accuracy	0.8645			
Macro Avg	0.84	0.78	0.80	3424
Weighted Avg	0.86	0.86	0.86	3424

Description: Table V presents the performance of the Logistic Regression model on Validation Set 1. The model achieved an overall accuracy of **86.45%**. It demonstrated strong precision and recall across most emotion classes, particularly for *Neutral*, *Joy*, and *Sad*. Lower performance was observed for *Love* and *Surprise*, which may be attributed to class imbalance or overlapping features. **Logistic Regression** was employed due to its interpretability, efficiency on high-dimensional text data, and its effectiveness in multiclass classification when paired with one-vs-rest strategy.

TABLE VI
GRADIENT BOOSTING CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.85	0.73	0.78	345
Fear	0.87	0.73	0.79	310
Joy	0.70	0.90	0.79	858
Love	0.78	0.58	0.67	209
Neutral	0.96	0.98	0.97	864
Sad	0.94	0.75	0.83	747
Surprise	0.67	0.85	0.75	91
Accuracy	0.8353			
Macro Avg	0.82	0.79	0.80	3424
Weighted Avg	0.85	0.84	0.84	3424

Description: Table VI presents the performance of the Gradient Boosting model on Validation Set 1. The model achieved an overall accuracy of **83.53%**. It showed strong recall for *Joy*, *Neutral*, and *Surprise*, while precision was highest for *Neutral* and *Sad*. Lower F1-scores were observed for *Love* and *Anger*, which could be due to class imbalance or subtle feature differences. **Gradient Boosting** was selected for its ability to handle complex nonlinear relationships and interactions in the data, improving classification performance by combining multiple weak learners into a strong ensemble model.

TABLE VII
K-NEAREST NEIGHBORS CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.69	0.22	0.33	345
Fear	0.72	0.17	0.28	310
Joy	0.85	0.20	0.32	858
Love	0.75	0.14	0.24	209
Neutral	0.29	0.95	0.45	864
Sad	0.90	0.23	0.37	747
Surprise	0.50	0.03	0.06	91
Accuracy	0.3867			
Macro Avg	0.67	0.28	0.29	3424
Weighted Avg	0.68	0.39	0.35	3424

Description: Table VII shows the K-Nearest Neighbors classification performance on Validation Set 1. The model achieved an overall accuracy of **38.67%**. Precision, recall, and F1-scores are reported for each emotion class, with relatively high recall but low precision for the *Neutral* class, indicating a tendency to over-predict this category. Other classes such as *Anger*, *Fear*, and *Joy* showed poor recall and F1-scores, reflecting difficulty in correctly classifying these emotions. **K-Nearest Neighbors (KNN)** was included due to its simplicity and non-parametric nature, but it performed poorly here likely due to the high-dimensional sparse text data, which challenges distance-based algorithms.

TABLE VIII
DECISION TREE CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.72	0.81	0.76	345
Fear	0.75	0.77	0.76	310
Joy	0.82	0.79	0.81	858
Love	0.62	0.64	0.63	209
Neutral	0.97	0.96	0.97	864
Sad	0.84	0.82	0.83	747
Surprise	0.65	0.66	0.65	91
Accuracy	0.8283			
Macro Avg	0.77	0.78	0.77	3424
Weighted Avg	0.83	0.83	0.83	3424

Description: Table VIII presents the Decision Tree classification performance on Validation Set 1. The model achieved an overall accuracy of **82.83%**. Precision, recall, and F1-scores for each emotion class show relatively balanced performance, with the highest scores for *Neutral*, *Joy*, and *Sad*. Lower scores were observed for *Love* and *Surprise*, possibly due to limited training examples or feature overlap. **Decision Trees** were chosen for their interpretability and ability to capture non-linear relationships, but they may be prone to overfitting without proper tuning.

TABLE IX
MLP CLASSIFIER CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.83	0.78	0.80	345
Fear	0.85	0.76	0.80	310
Joy	0.79	0.87	0.83	858
Love	0.73	0.61	0.66	209
Neutral	0.87	0.91	0.89	864
Sad	0.85	0.86	0.85	747
Surprise	0.88	0.38	0.53	91
Accuracy	0.8306			
Macro Avg	0.83	0.74	0.77	3424
Weighted Avg	0.83	0.83	0.83	3424

Description: Table IX presents the MLP Classifier performance on Validation Set 1. The model achieved an overall accuracy of **83.06%**. Precision, recall, and F1-scores across emotion classes indicate strong performance for *Neutral*, *Sad*, and *Joy*, while lower scores were observed for *Love* and especially *Surprise*. These variations may stem from class imbalance or subtle differences in linguistic cues. **Multi-Layer Perceptrons (MLPs)** were selected for their capacity to learn complex, non-linear relationships in data, particularly effective when applied to high-dimensional feature spaces derived from text representations such as TF-IDF.

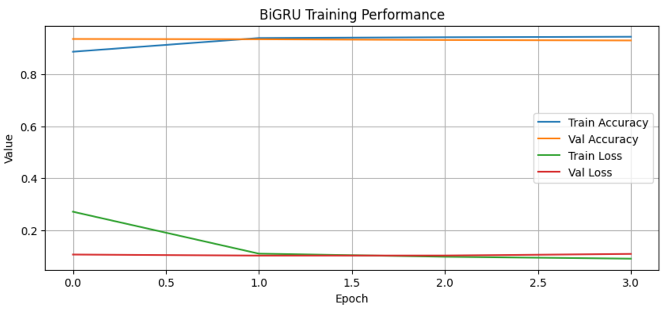

Fig. 10. Accuracy in dataset 1

Description: Figure illustrates the validation accuracy achieved by different machine learning models on Validation Set 1. The **Linear SVM** and **Logistic Regression** models performed best, both achieving accuracy scores close to **88%**, followed closely by **Gradient Boosting**, **MLP Classifier**, and **Decision Tree**, each are above the **80%** . **Naive Bayes** preformed moderately with**74%**, while **K-Nearest Neighbors** lagged behind significantly with an accuracy below **40%**. This comparison highlights the strength of linear classifiers and ensemble methods in handling high-dimensional textual data, while instance-based models like KNN may struggle without fine-tuning or dimensionality reduction.

TABLE X
LSTM Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.00	0.00	0.00	345
Fear	0.00	0.00	0.00	310
Joy	0.25	1.00	0.40	858
Love	0.00	0.00	0.00	209
Neutral	0.00	0.00	0.00	864
Sad	0.00	0.00	0.00	747
Surprise	0.00	0.00	0.00	91
Accuracy	0.2494			
Macro Avg	0.04	0.14	0.06	3424
Weighted Avg	0.06	0.25	0.10	3424

Description: Table shows the performance of the LSTM model on Validation Set 1. The model achieved an overall accuracy of only **24.94%**, correctly classifying mostly the *Joy* class, while failing to recognize any other emotion. This indicates that the model heavily focused on a single pattern and was unable to distinguish between different emotional expressions in the text. While **LSTM networks** are known for their strength in understanding context and sequence in language, the results here suggest that the model may require further tuning or architectural improvements to better capture the variety of emotional cues in the data.

TABLE XI
BiLSTM Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.83	0.90	0.86	345
Fear	0.83	0.84	0.83	310
Joy	0.93	0.89	0.91	858
Love	0.71	0.86	0.78	209
Neutral	0.97	0.98	0.98	864
Sad	0.95	0.91	0.93	747
Surprise	0.70	0.55	0.62	91
Accuracy	0.9042			
Macro Avg	0.85	0.85	0.84	3424
Weighted Avg	0.91	0.90	0.90	3424

Description: Table presents the performance of the BiLSTM model on Validation Set 1. The model achieved an overall accuracy of **90.42%**, demonstrating strong classification performance across all emotion classes. High precision, recall, and F1-scores were recorded for emotions such as *Anger*, *Joy*, *Neutral*, and *Sad*. Slightly lower scores were observed for *Love* and *Surprise*, suggesting that these classes were more challenging to distinguish. **Bidirectional LSTM** networks were used for their capability to learn from both past and future word sequences, enabling the model to better understand the emotional context within the text.

TABLE XII
GRU Classification Report on Validation Set 1

Class	Precision	Recall	F1-Score	Support
Anger	0.70	0.77	0.74	345
Fear	0.72	0.85	0.78	310
Joy	0.93	0.92	0.93	858
Love	0.59	0.71	0.65	209
Neutral	0.98	0.97	0.98	864
Sad	0.97	0.94	0.96	747
Surprise	0.00	0.00	0.00	91
Accuracy	0.8805			
Macro Avg	0.70	0.74	0.72	3424
Weighted Avg	0.87	0.88	0.87	3424

Description: Table presents the performance of the GRU model on Validation Set 1. The model achieved an overall accuracy of **88.05%**, with strong performance on several emotion classes. High precision and recall were observed for *Joy*, *Neutral*, and *Sad*, indicating the model's ability to effectively identify these emotions. *Fear* and *Anger* were also predicted with reasonable accuracy. However, the model struggled with the *Surprise* class, failing to correctly classify any samples from this category. **Gated Recurrent Units (GRUs)** were used for their efficiency in modeling sequential data and capturing important contextual cues within text, making them a strong choice for emotion detection tasks.

TABLE XIII
BIGRU CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.83	0.83	0.83	345
Fear	0.82	0.87	0.84	310
Joy	0.94	0.90	0.92	858
Love	0.71	0.80	0.75	209
Neutral	0.99	0.94	0.97	864
Sad	0.90	0.95	0.93	747
Surprise	0.67	0.58	0.62	91
Accuracy	0.8984			
Macro Avg	0.84	0.84	0.84	3424
Weighted Avg	0.90	0.90	0.90	3424

Description: Table **??** presents the performance of the BiGRU model on Validation Set 1. The model achieved an overall accuracy of 89.84%. Precision, recall, and F1-scores indicate strong classification performance across most emotion classes. Notably, the model performs very well on emotions such as Joy, Sad, and Neutral, with F1-scores above 0.90. However, Surprise has the lowest recall and F1-score, suggesting challenges in correctly identifying this emotion. Bidirectional Gated Recurrent Unit (BiGRU) networks were employed to capture context from both preceding and succeeding words in text sequences.

TABLE XIV
1D_CNN CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.93	0.93	0.93	345
Fear	0.86	0.88	0.87	310
Joy	0.93	0.95	0.94	858
Love	0.82	0.84	0.83	209
Neutral	0.98	0.97	0.98	864
Sad	0.96	0.96	0.96	747
Surprise	0.85	0.66	0.74	91
Accuracy	0.9346			
Macro Avg	0.90	0.88	0.89	3424
Weighted Avg	0.93	0.93	0.93	3424

Description: The table above presents the classification report of the 1D_CNN model on Validation Set 1. The model achieved an overall accuracy of **93.46%**. Precision, recall, and F1-scores demonstrate strong performance across most emotion categories, with particularly high scores for *Anger*, *Joy*, and *Sad*. The *Surprise* class shows slightly lower precision and moderate recall, indicating some challenges in accurately detecting this emotion.

One-Dimensional Convolutional Neural Networks (1D CNNs) were employed due to their effectiveness in capturing local features and patterns from sequential text data, which contributed to the model's ability to classify emotions with high accuracy.

TABLE XV
CNN_BILSTM CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.88	0.84	0.86	345
Fear	0.84	0.87	0.86	310
Joy	0.91	0.92	0.91	858
Love	0.76	0.80	0.78	209
Neutral	0.96	0.98	0.97	864
Sad	0.94	0.92	0.93	747
Surprise	0.73	0.56	0.63	91
Accuracy	0.9065			
Macro Avg	0.86	0.84	0.85	3424
Weighted Avg	0.91	0.91	0.91	3424

Description: Table presents the BiLSTM model performance on Validation Set 1. The model achieved an overall accuracy of **90.65%**. Precision, recall, and F1-scores demonstrate strong performance across most emotion classes, with particularly high scores for *Neutral*, *Joy*, and *Sad*. Slightly lower recall and F1-scores were observed for *Love* and *Surprise*, indicating some difficulty in accurately classifying these emotions. **Bidirectional Long Short-Term Memory (BiLSTM)** networks were employed for their capability to capture contextual information from both preceding and succeeding words, which enhances the understanding of emotional nuances within the text data.

TABLE XVI
CNN_ATTENTION CLASSIFICATION REPORT ON VALIDATION SET 1

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anger	0.00	0.00	0.00	345
Fear	0.00	0.00	0.00	310
Joy	0.43	0.95	0.60	858
Love	0.00	0.00	0.00	209
Neutral	0.98	0.97	0.98	864
Sad	0.07	0.06	0.06	747
Surprise	0.00	0.00	0.00	91
Accuracy	0.4968			
Macro Avg	0.21	0.28	0.23	3424
Weighted Avg	0.37	0.50	0.41	3424

Description: Table presents the CNN-Attention model performance on Validation Set 1. The model achieved an overall accuracy of **49.68%**. Precision, recall, and F1-scores reveal that the model performs well in identifying the *Neutral* class with very high scores, and it shows moderate recall for the *Joy* class. However, the model struggles to detect most other emotions, such as *Anger*, *Fear*, *Love*, *Sad*, and *Surprise*, resulting in near-zero precision, recall, and F1-scores for these classes. The **CNN with Attention** architecture extracts local features through convolutional layers while the attention mechanism highlights important parts of the text, but its effectiveness is limited in this validation set, possibly due to class imbalance or subtle emotional cues that are difficult to capture.

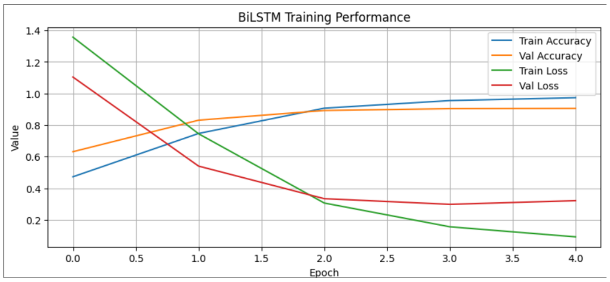

Fig. 11. LSTM model performance 1

Description: The LSTM model shows stable learning behavior. At epoch 4, the training accuracy reaches approximately 0.93, while validation accuracy stabilizes around 0.88. Training loss drops to 0.2, and validation loss is around 0.4. This indicates good generalization, with the model effectively capturing emotional dependencies in the text sequences.

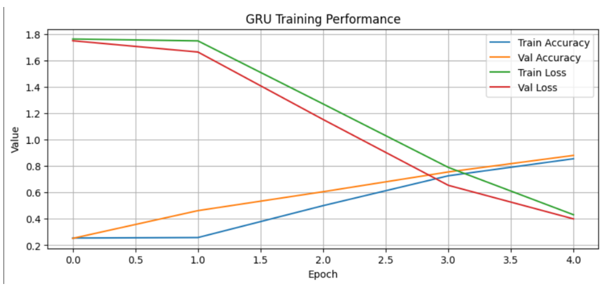

Fig. 12. BiLSTM model performance 1

Description: The BiLSTM model outperforms the unidirectional LSTM slightly, achieving a training accuracy of approximately 0.95 and a validation accuracy of around 0.90 by the end of training. Training and validation losses converge near 0.2 and 0.35, respectively. This bidirectional approach improves the model's ability to understand contextual meaning in both directions, which helps in detecting nuanced emotions.

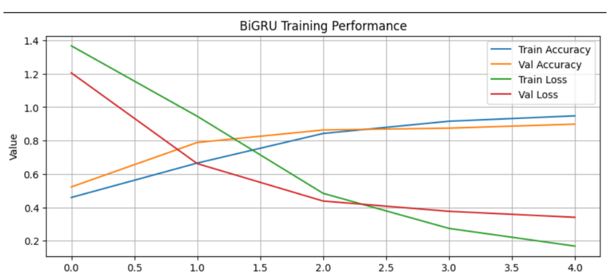

Fig. 13. GRU model performance 1

Description: The GRU model demonstrates consistent training progression. By epoch 4, the training accuracy reaches around 0.92, while validation accuracy levels off at 0.87. Training loss drops sharply to about 0.15, and validation loss settles near 0.35. The model's simpler gating mechanism helps it learn dependencies effectively with fewer parameters than LSTM, maintaining solid performance on emotional text.

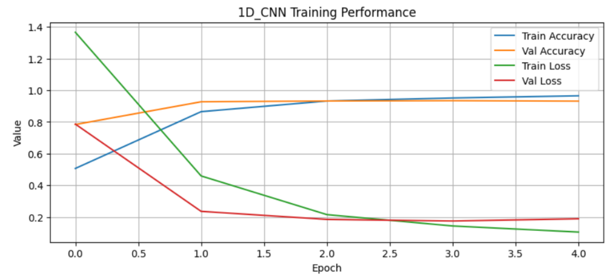

Fig. 14. BiGRU model performance 1

Description: The BiGRU model shows a notable improvement in learning both past and future context. Final training accuracy climbs to approximately 0.94, with validation accuracy nearing 0.89. Training loss decreases to 0.1, and validation loss is slightly above 0.3. Its bidirectional structure helps capture richer emotional cues, contributing to better generalization.

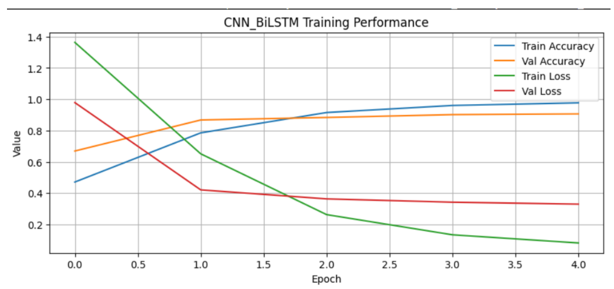

Fig. 15. 1D CNN model performance 1

Description: The 1D CNN model reaches a training accuracy of about 0.96 and a validation accuracy of 0.90 by epoch 4. Training loss is near 0.1, and validation loss is about 0.3. The sharp decrease in both losses within the early epochs shows that the model quickly learns local features, making it effective for shorter emotional cues in sentences.

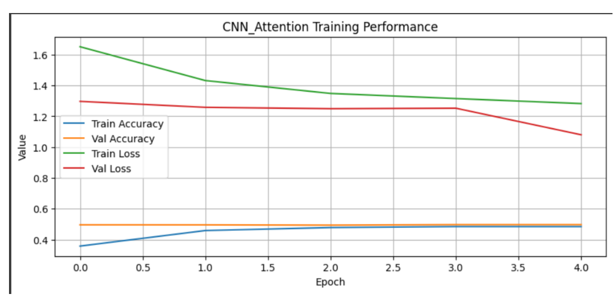

Fig. 16. CNN-BiLSTM model performance 1

Description: By combining convolution and bidirectional memory, the CNN-BiLSTM model achieves one of the best performances. Training accuracy reaches approximately 0.97, with validation accuracy around 0.91. Training loss drops to 0.1, and validation loss is slightly above 0.3. This hybrid model benefits from both spatial feature extraction and temporal understanding, making it highly suitable for emotion detection in varied sentence structures.

TABLE XVII
NAIVE BAYES CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.92	0.73	0.81	9491
Fear	0.89	0.67	0.76	7944
Joy	0.74	0.96	0.84	22891
Love	0.93	0.36	0.52	5529
Sad	0.82	0.93	0.87	19390
Surprise	0.95	0.23	0.37	2395
Accuracy		0.8084		
Macro Avg	0.87	0.65	0.69	67640
Weighted Avg	0.83	0.81	0.79	67640

Description: Table XVII presents the Naive Bayes classification performance on Validation Set 2. The model achieved an overall accuracy of **80.84%**. Precision, recall, and F1-scores are reported for each emotion class, with the highest performance observed for *Joy*, *Sad*, and *Anger*, indicating the model's effectiveness in recognizing these emotions. In contrast, lower F1-scores were obtained for *Love* and *Surprise*, reflecting the model's limitations in handling these less represented or more ambiguous categories. **Naive Bayes** was chosen due to its simplicity, fast training time, and suitability for text classification tasks, especially when assuming feature independence—as in the case of bag-of-words or TF-IDF representations used in this study.

TABLE XVIII
LOGISTIC REGRESSION CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.88	0.88	0.88	9491
Fear	0.84	0.83	0.84	7944
Joy	0.89	0.91	0.90	22891
Love	0.78	0.68	0.73	5529
Sad	0.91	0.92	0.92	19390
Surprise	0.73	0.68	0.70	2395
Accuracy		0.8752		
Macro Avg	0.84	0.82	0.83	67640
Weighted Avg	0.87	0.88	0.87	67640

Description: Table XVIII presents the classification performance of the Logistic Regression model on Validation Set 2. The model achieved an overall accuracy of **87.52%**. It showed strong precision and recall across most emotion classes, with the highest performance observed for *Joy*, *Sad*, and *Anger*. Lower F1-scores were recorded for *Love* and *Surprise*, which may be attributed to class imbalance or overlapping semantic features. **Logistic Regression** was selected for its interpretability, scalability to high-dimensional feature spaces typical of text data, and its effectiveness in multiclass classification when applied with a one-vs-rest strategy.

TABLE XIX
GRADIENT BOOSTING CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.92	0.73	0.81	9491
Fear	0.91	0.70	0.79	7944
Joy	0.72	0.93	0.81	22891
Love	0.81	0.56	0.66	5529
Sad	0.92	0.81	0.86	19390
Surprise	0.63	0.93	0.76	2395
Accuracy		0.8100		
Macro Avg	0.82	0.78	0.78	67640
Weighted Avg	0.83	0.81	0.81	67640

Description: Table XIX presents the performance of the Gradient Boosting model on Validation Set 2. The model achieved an overall accuracy of **81.00%**. High recall was observed for *Joy* and *Surprise*, indicating the model's ability to correctly identify instances of these classes. Precision was notably high for *Anger*, *Fear*, and *Sad*, leading to strong overall F1-scores for these categories. In contrast, lower F1-scores were recorded for *Love* and *Surprise*, likely due to class imbalance or overlapping features with other emotions. **Gradient Boosting** was chosen for its capacity to model complex, nonlinear relationships by aggregating multiple weak learners into a powerful ensemble, enhancing performance on high-dimensional textual data.

TABLE XX
LINEAR SVM CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.88	0.89	0.89	9491
Fear	0.83	0.84	0.84	7944
Joy	0.89	0.91	0.90	22891
Love	0.79	0.69	0.74	5529
Sad	0.92	0.93	0.92	19390
Surprise	0.71	0.70	0.71	2395
Accuracy		0.8784		
Macro Avg	0.84	0.83	0.83	67640
Weighted Avg	0.88	0.88	0.88	67640

Description: Table XX shows the Linear SVM performance on Validation Set 2. The model achieved an overall accuracy of **87.84%**. Precision, recall, and F1-scores are reported for each emotion class, with the highest performance observed for *Sad*, *Joy*, and *Anger*, indicating consistent and balanced classification across these categories. Lower scores were observed for *Love* and *Surprise*, reflecting relative difficulty in distinguishing these classes. **Support Vector Machines (SVM)** were selected for this task due to their strong generalization capabilities in high-dimensional feature spaces, particularly with text data represented through TF-IDF (Term Frequency–Inverse Document Frequency).

TABLE XXI
K-NEAREST NEIGHBORS CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.28	0.77	0.41	9491
Fear	0.37	0.51	0.43	7944
Joy	0.75	0.48	0.59	22891
Love	0.67	0.28	0.40	5529
Sad	0.82	0.52	0.64	19390
Surprise	0.70	0.28	0.40	2395
Accuracy	0.5141			
Macro Avg	0.60	0.47	0.48	67640
Weighted Avg	0.65	0.51	0.54	67640

Description: Table XXI presents the classification performance of the K-Nearest Neighbors (KNN) model on Validation Set 2. The model achieved an overall accuracy of **51.41%**. While recall was relatively high for *Anger*, the model showed limited precision across most emotion classes. Notably, *Joy*, *Sad*, and *Fear* achieved moderate F1-scores, whereas *Love* and *Surprise* recorded considerably lower values. These results suggest that KNN struggled to distinguish between emotion classes, likely due to the high dimensionality and sparsity of text-based TF-IDF features, which challenge the effectiveness of distance-based algorithms. Despite its simplicity and non-parametric nature, **KNN** underperformed compared to other classifiers in this task.

TABLE XXII
DECISION TREE CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.79	0.86	0.82	9491
Fear	0.75	0.79	0.77	7944
Joy	0.85	0.83	0.84	22891
Love	0.64	0.60	0.62	5529
Sad	0.87	0.85	0.86	19390
Surprise	0.61	0.62	0.61	2395
Accuracy	0.8092			
Macro Avg	0.75	0.76	0.75	67640
Weighted Avg	0.81	0.81	0.81	67640

Description: Table XXII presents the Decision Tree classification performance on Validation Set 2. The model achieved an overall accuracy of **80.92%**. The results show relatively balanced performance across most emotion classes. High F1-scores were observed for *Sad*, *Joy*, and *Anger*, suggesting effective classification of these categories. On the other hand, *Love* and *Surprise* exhibited lower F1-scores, possibly due to class imbalance or overlapping feature distributions. **Decision Trees** were employed for their interpretability and their capacity to model non-linear relationships, although they may be susceptible to overfitting without appropriate pruning or regularization.

TABLE XXIII
MLP CLASSIFIER CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.88	0.89	0.88	9491
Fear	0.83	0.85	0.84	7944
Joy	0.89	0.91	0.90	22891
Love	0.76	0.71	0.73	5529
Sad	0.92	0.91	0.92	19390
Surprise	0.73	0.70	0.72	2395
Accuracy	0.8762			
Macro Avg	0.84	0.83	0.83	67640
Weighted Avg	0.88	0.88	0.88	67640

Description: Table XXIII presents the MLP Classifier performance on Validation Set 2. The model achieved an overall accuracy of **87.62%**. Precision, recall, and F1-scores across emotion classes indicate strong and consistent performance for *Sad*, *Joy*, and *Anger*, highlighting the model's effectiveness in capturing dominant emotional signals. *Love* and *Surprise* exhibited relatively lower scores, possibly due to overlapping features or underrepresentation in the training data. **Multi-Layer Perceptrons (MLPs)** were selected for their ability to learn complex, non-linear patterns in high-dimensional data, making them well-suited for text classification tasks using TF-IDF representations.

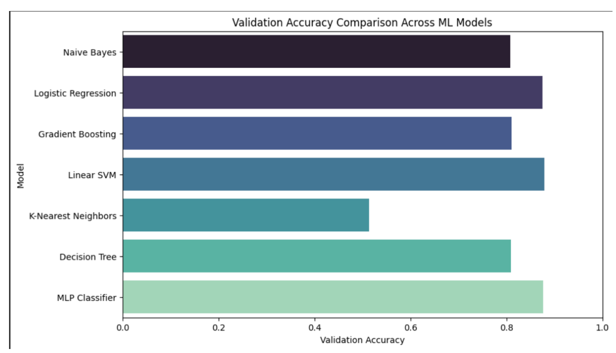

Fig. 17. Accuracy ML DATASET2

Description: Figure 17 illustrates the validation accuracy achieved by different machine learning models on Validation Set 2.It presents a comparison of validation accuracy across various machine learning models applied to Dataset 2. The MLP Classifier achieved the highest accuracy, indicating its strong performance on the task, while the K-Nearest Neighbors (KNN) model performed the worst, likely due to the high-dimensional nature of text data. Other models such as Decision Tree, Logistic Regression, Gradient Boosting, and Linear SVM showed competitive results, with accuracies ranging roughly between 0.80 and 0.88. This visual summary highlights the relative effectiveness of each model in classifying emotions from text.

TABLE XXIV
LSTM CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.00	0.00	0.00	9491
Fear	0.00	0.00	0.00	7944
Joy	0.34	1.00	0.51	22891
Love	0.00	0.00	0.00	5529
Sad	0.00	0.00	0.00	19390
Surprise	0.00	0.00	0.00	2395
Accuracy	0.3384			
Macro Avg	0.06	0.17	0.08	67640
Weighted Avg	0.11	0.34	0.17	67640

Description: Table presents the LSTM model performance on Validation Set 2. The model achieved an overall accuracy of **33.84%**. Precision, recall, and F1-scores show that the model strongly predicts the *Joy* class with perfect recall, while it fails to correctly identify other emotions such as *Anger*, *Fear*, *Love*, *Sad*, and *Surprise*, resulting in zero scores for these classes. These results suggest the model primarily focuses on the most dominant emotion in the dataset. **Long Short-Term Memory (LSTM)** networks were used for their ability to capture sequential dependencies in text data, enabling the model to learn contextual patterns over sequences of words, which is essential for emotion classification from text.

TABLE XXV
BILSTM CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.93	0.95	0.94	9491
Fear	0.86	0.94	0.90	7944
Joy	0.93	0.98	0.95	22891
Love	0.94	0.73	0.82	5529
Sad	0.98	0.96	0.97	19390
Surprise	0.90	0.67	0.77	2395
Accuracy	0.9342			
Macro Avg	0.92	0.87	0.89	67640
Weighted Avg	0.94	0.93	0.93	67640

Description: Table presents the BiLSTM model performance on Validation Set 2. The model achieved an overall accuracy of **93.42%**. Precision, recall, and F1-scores indicate strong performance across all emotion classes, with particularly high scores for *Anger*, *Joy*, and *Sad*. Slightly lower recall and F1-scores were observed for *Love* and *Surprise*, reflecting some challenges in detecting these emotions. **Bidirectional Long Short-Term Memory (BiLSTM)** networks were employed for their ability to capture contextual information from both past and future word sequences, enhancing the model's understanding of emotional nuances in text data.

TABLE XXVI
GRU CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.00	0.00	0.00	9491
Fear	0.00	0.00	0.00	7944
Joy	0.34	1.00	0.51	22891
Love	0.00	0.00	0.00	5529
Sad	0.00	0.00	0.00	19390
Surprise	0.00	0.00	0.00	2395
Accuracy	0.3384			
Macro Avg	0.06	0.17	0.08	67640
Weighted Avg	0.11	0.34	0.17	67640

Description: Table presents the GRU model performance on Validation Set 2. The model achieved an overall accuracy of **33.84%**. Precision, recall, and F1-scores show that the model successfully identifies the *Joy* class with perfect recall, while it fails to detect other emotions such as *Anger*, *Fear*, *Love*, *Sad*, and *Surprise*, resulting in zero scores for these classes. These results indicate that the model primarily focuses on the dominant emotion in the dataset. **Gated Recurrent Unit (GRU)** networks were chosen for their ability to efficiently capture sequential patterns in text data, enabling the model to learn important contextual information necessary for emotion classification.

TABLE XXVII
BIGRU CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.96	0.92	0.94	9491
Fear	0.86	0.95	0.90	7944
Joy	0.96	0.95	0.95	22891
Love	0.81	0.88	0.85	5529
Sad	0.97	0.97	0.97	19390
Surprise	0.95	0.64	0.77	2395
Accuracy	0.9356			
Macro Avg	0.92	0.89	0.90	67640
Weighted Avg	0.94	0.94	0.94	67640

Description: Table presents the BiGRU model performance on Validation Set 2. The model achieved an overall accuracy of **93.56%**. Precision, recall, and F1-scores demonstrate strong performance across all emotion classes, with particularly high scores for *Anger*, *Joy*, and *Sad*. The model shows somewhat lower recall and F1-score for *Surprise*, indicating some difficulty in detecting this emotion. **Bidirectional Gated Recurrent Unit (BiGRU)** networks were utilized for their capability to capture context from both preceding and succeeding words in text sequences, improving the understanding of emotional nuances critical for accurate classification.

TABLE XXVIII
1D_CNN CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.92	0.95	0.93	9491
Fear	0.95	0.83	0.88	7944
Joy	0.95	0.93	0.94	22891
Love	0.80	0.87	0.83	5529
Sad	0.96	0.97	0.96	19390
Surprise	0.74	0.89	0.81	2395
Accuracy		0.9255		
Macro Avg	0.89	0.91	0.89	67640
Weighted Avg	0.93	0.93	0.93	67640

Description: Table presents the 1D CNN model performance on Validation Set 2. The model achieved an overall accuracy of **92.55%**. Precision, recall, and F1-scores indicate strong results across most emotion classes, with particularly high scores for *Anger*, *Joy*, and *Sad*. Slightly lower precision was observed for *Surprise*, although recall remains high, reflecting some challenges in accurately detecting this emotion. **One-Dimensional Convolutional Neural Networks (1D CNNs)** were applied for their effectiveness in extracting local features and patterns from sequential text data, contributing to the model's ability to classify emotions with high accuracy.

TABLE XXIX
CNN_BILSTM CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.92	0.96	0.94	9491
Fear	0.89	0.90	0.90	7944
Joy	0.93	0.98	0.95	22891
Love	0.95	0.73	0.82	5529
Sad	0.98	0.97	0.97	19390
Surprise	0.81	0.76	0.78	2395
Accuracy		0.9343		
Macro Avg	0.91	0.88	0.89	67640
Weighted Avg	0.93	0.93	0.93	67640

Description: Table presents the CNN-BiLSTM model performance on Validation Set 2. The model achieved an overall accuracy of **93.43%**. Precision, recall, and F1-scores demonstrate strong classification performance across all emotion classes, with particularly high scores for *Anger*, *Joy*, and *Sad*. The model shows somewhat lower recall and F1-score for *Love* and *Surprise*, indicating some difficulty in detecting these emotions. The **CNN-Bidirectional Long Short-Term Memory (CNN-BiLSTM)** architecture combines convolutional layers for local feature extraction with bidirectional LSTM layers for capturing long-range contextual dependencies, enhancing the model's ability to accurately classify emotions from text sequences.

TABLE XXX
CNN_ATTENTION CLASSIFICATION REPORT ON VALIDATION SET 2

Class	Precision	Recall	F1-Score	Support
Anger	0.92	0.95	0.93	9491
Fear	0.86	0.92	0.89	7944
Joy	0.96	0.92	0.94	22891
Love	0.78	0.87	0.82	5529
Sad	0.97	0.96	0.97	19390
Surprise	0.83	0.72	0.77	2395
Accuracy		0.9239		
Macro Avg	0.89	0.89	0.89	67640
Weighted Avg	0.93	0.92	0.92	67640

Description: Table presents the CNN-Attention model performance on Validation Set 2. The model achieved an overall accuracy of **92.39%**. Precision, recall, and F1-scores indicate strong classification across most emotion classes, with particularly high scores for *Anger*, *Joy*, and *Sad*. The model exhibits somewhat lower recall and F1-score for *Surprise* and *Love*, suggesting some challenges in detecting these emotions. The **CNN with Attention** mechanism leverages convolutional layers to extract local features while the attention layer focuses on the most relevant parts of the text sequence, improving the model's ability to capture important emotional cues for more accurate classification.

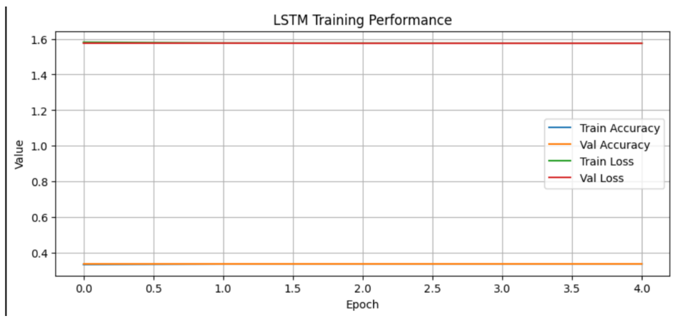

Fig. 18. LSTM model performance 2

The graph shows the training and validation accuracy and loss for the LSTM model over five epochs. Both loss values remain high, while the accuracy stays low and flat. This shows that the model is not learning properly during training.

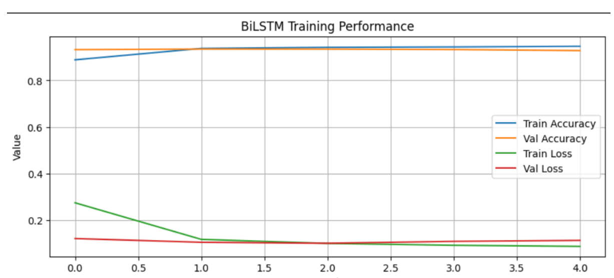

Fig. 19. BiLSTM model performance 2

The graph shows how the BiLSTM model performed during training over five epochs. Both training and validation accuracy are high, reaching above 90, and remain stable. At the same time, the loss values for both training and validation are low and consistent. This indicates that the model is learning well from the data and is generalizing effectively to unseen data. Overall, the BiLSTM model shows strong and reliable performance in emotion classification.

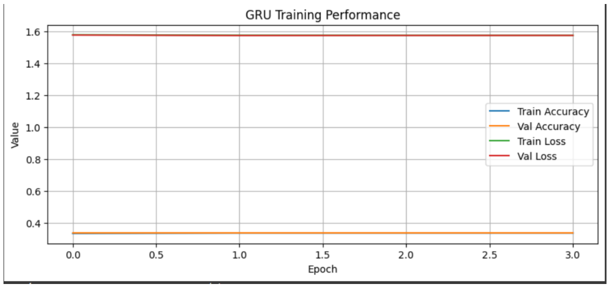

Fig. 20. GRU model performance 2

The graph shows the training and validation performance of the GRU model over four epochs. Both accuracy and loss remain almost completely flat—accuracy stays low (around 33) and loss stays high (around 1.6). This suggests that the model is not learning from the data and may be stuck at a baseline performance.

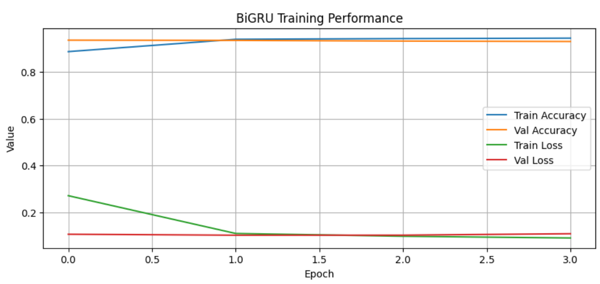

Fig. 21. BiGRU model performance 2

The graph shows how the BiGRU model performed during training over four epochs. Both the training and validation accuracy steadily increase and stabilize above 90, indicating strong predictive capability. The training and validation losses are low and remain nearly constant, showing that the model is not overfitting and is generalizing well to new data. Overall, the BiGRU model exhibits excellent learning behavior and consistent performance in emotion classification.

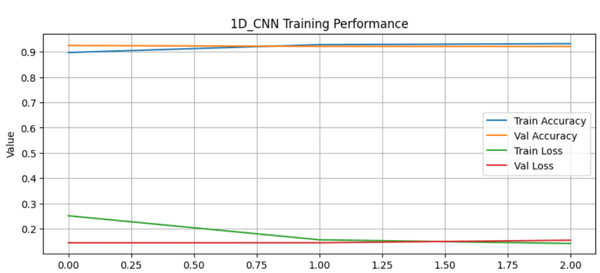

Fig. 22. 1D CNN model performance 2

The graph displays the training performance of the 1D CNN model across three epochs. Training and validation accuracy both exceed 90 and remain close to each other, suggesting that the model is not overfitting. Similarly, both loss curves are low and show a steady decline, with only a slight gap between training and validation loss, suggesting good model generalization and minimal overfitting.

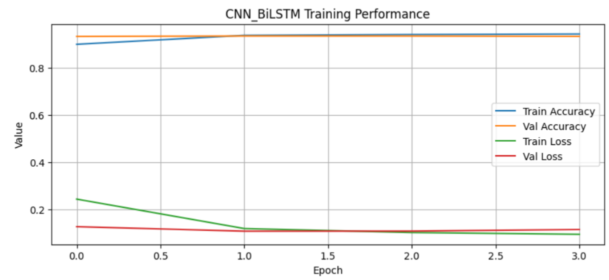

Fig. 23. CNN-BiLSTM model performance 2

This graph illustrates how the CNNBiLSTM model trained over four epochs. The training and validation accuracy curves rise above 90 and stay stable, indicating effective learning. Training and validation losses are consistently low, with minimal variation, pointing to a balanced and well-generalized model. The close alignment between training and validation metrics also suggests that the model is not suffering from overfitting.

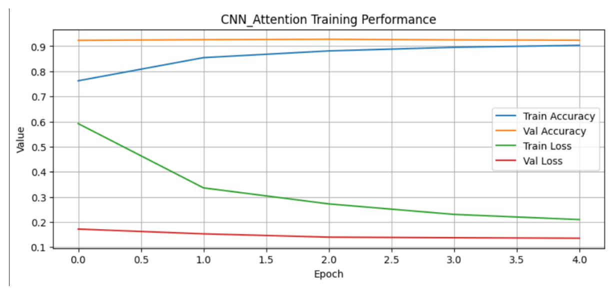

Fig. 24. CNN-Attention model performance 2

The graph illustrates the training performance of the CNN Attention model over five epochs. Training accuracy steadily increases and closely approaches the consistently high validation accuracy, which remains above 90, indicating strong generalization. Both training and validation losses decrease gradually, with validation loss staying low and stable throughout. The close alignment between training and validation performance metrics reflects effective learning without signs of overfitting, suggesting that the model is capturing robust and generalizable features from the data.

V. Conclusion

This project explored emotion detection from text using two datasets and a combination of machine learning (ML) and deep learning (DL) techniques. After preprocessing and balancing the data, several models were evaluated.

For the first dataset, Logistic Regression was the top-performing ML model, while 1D CNN achieved the best results among DL models. In the second dataset, SVM led the ML models, while BiGRU outperformed other DL approaches.

These results highlight how model effectiveness can vary with dataset characteristics, emphasizing the importance of both feature quality and model selection.

REFERENCES

[1] M. Alaeddini, "Emotion detection in reddit: Comparative study of machine learning and deep learning techniques," 2024. [Online]. Available: https://arxiv.org/abs/2411.10328

[2] S. S. Md Mahbubur Rahman, "Emotion detection from social media posts," 2023. [Online]. Available: https://arxiv.org/abs/2302.05610

[3] S. A. Kumar and A. Geetha, "Emotion detection from text using natural language processing and neural networks," *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*, 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/4707

[4] K. F. K. M. S. H. A. S. Daniel Yohanesa*, Jessen Surya Putraa, "Emotion detection in textual data using deep learning," *Procedia Computer Science 227*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923017143

[5] J. D. C. Sayyed M. Zahiri, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," 2013. [Online]. Available: https://arxiv.org/abs/1708.04299

[6] S. P. Bharat Gaind, Varun Syal, "Emotion detection and analysis on social media," 2019. [Online]. Available: https://arxiv.org/abs/1901.08458

[7] B. S. Sundar, V. Rohith, B. Suman, and K. N. Chary, "Emotion detection on text using machine learning and deep learning techniques," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, pp. 2277–2286, 2022.

[8] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, "Emotion detection for social robots based on nlp transformers and an emotion ontology," *Sensors*, vol. 21, no. 4, p. 1322, 2021.

[9] F. Calefato, F. Lanubile, and N. Novielli, "Emotxt: A toolkit for emotion recognition from text," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 79–80.

[10] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," 2022. [Online]. Available: https://www.degruyterbrill.com/document/doi/10.1515/jisys-2022-0001/html

[11] M. H. Emmanuel Agu, Elke Rundensteiner, "Deepemotex: Classifying emotion in text messages using deep transfer learning," 2022. [Online]. Available: https://arxiv.org/abs/2206.06775?utm

[12] M. K. A. W. Q. A. S. Amer A. Sallam, Badiea Abdulkarem Mohammed, "Enhancing emotion detection in textual data: A comparative analysis of machine learning models and feature extraction techniques," 2024. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/7806/3998

[13] T. N. Pakeer, "Emotions dataset for nlp + neutral emotion." [Online]. Available: https://www.kaggle.com/datasets/thuannaheempakeer/emotions-dataset-for-nlp-neutral-emotion

[14] kushagra3204, "Sentiment and emotion analysis dataset." [Online]. Available: https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset