Text Emotion Analysis

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Computer Science Engineering

by

Himanshu Gautam - Roll No. 21UCS240
Anurag Singh - Roll No. 21UCC020
Mohit Jain - Roll No. 21UCS131

Under Guidance of
Dr. Durga Prasad Mishra
Dr. Navneet garg

LNMIIT
The LNM Institute of
Information Technology

Department of Computer Science Engineering
The LNM Institute of Information Technology, Jaipur

November 2024

©The LNM INSTITUTE OF INFORMATION TECHNOLOGY Jaipur-2024.

All rights reserved.

The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled “Text Emotion Analysis” , submitted by Himanshu Gautam (Roll no 21ucs240), Anurag Singh (Roll no 21UCC020) and Mohit Jain (Roll no 21UCS131) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by them at the Department of Computer Science Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2023-2024 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this report is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

Date

Adviser: Name of BTP Supervisor

Acknowledgments

We extend our heartfelt gratitude to Dr. Durga Prasad Mishra and Dr. Navneet Garg for their invaluable guidance, unwavering support, and insightful feedback throughout the duration of our BTech project on ”Text Emotion Analysis”.

Their combined expertise, encouragement, and dedication have been instrumental in shaping our research into a comprehensive study. Their constructive criticism and encouragement propelled us to strive for excellence and explore new horizons in this fascinating field.

We are deeply thankful to them for generously sharing their knowledge, wisdom, and resources, which significantly enriched this project and enhanced our understanding of the subject matter.

Abstract

Emotional recognition has emerged as a crucial area of research, offering valuable insights across various domains, including psychotherapy, customer experience, and human-computer interaction. Emotions manifest through speech, facial expressions, gestures, and written text. This study focuses on detecting emotions in textual data, a challenge rooted in Natural Language Processing (NLP) and deep learning. Leveraging semantic text analysis, word embeddings play a pivotal role in capturing the syntactic and semantic nuances of textual content, enhancing tasks like machine translation and sentiment analysis. We propose a deep learning-assisted semantic text analysis (DLSTA) model that achieved high performance metrics, including accuracy, recall, and F1 scores above 91%, outperforming baseline models.

Building upon prior advancements, this study delves into dimensional emotion recognition, which maps emotions onto continuous scales such as valence, arousal, and dominance (VAD), enabling a deeper and more precise understanding of emotional states. Leveraging a transformer-based model, fine-tuned using a novel dimensional emotion dataset comprising 75,503 samples, the approach demonstrates significant enhancements over state-of-the-art techniques. The model addresses critical challenges, including heterogeneity in annotations and domain transfer, achieving remarkable Pearson correlation coefficients of 0.90 for valence, 0.77 for arousal, and 0.64 for dominance. By capturing the nuanced and subjective facets of emotional expressions, this methodology complements traditional categorical models, advancing precision, robustness, and applicability in diverse settings. These improvements hold particular promise for applications in therapeutic contexts and interactive systems, where nuanced emotion detection can significantly enhance user experiences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Area of Work

Emotional recognition has emerged as an essential field of research, deeply intertwined with the study of human behavior and interactions. Emotions are fundamental to the human experience, often defined as "a complex pattern of reactions, including experiential, behavioral, and physiological elements." They influence decision-making, communication, and relationships. Understanding emotions expressed in text—whether through social media posts, product reviews, or other written formats—offers valuable insights for various domains, including marketing, healthcare, and human-computer interaction.

The advent of digital platforms has resulted in a surge of text-based content, providing a rich source of data for emotion analysis. Natural Language Processing (NLP), coupled with advancements in deep learning, has revolutionized how text is processed and understood. This research aims to delve into text emotion analysis using a combination of machine learning techniques, focusing on extracting meaningful emotional information from large-scale textual data. By using deep learning-assisted semantic text analysis (DLSTA) and integrating emotional dimensions such as valence, arousal, and dominance, this study aims to create precise and scalable systems for understanding and responding to human emotions.

1.2 Problem Addressed

The exponential growth of digital content in the form of social media posts, product reviews, news articles, and online conversations has led to an increasing need for accurate and efficient methods of understanding and analyzing human emotions expressed in text. However, the inherent complexities of language, including sarcasm, ambiguity, and cultural context, pose significant challenges to traditional approaches of emotion analysis. Consequently, there is a pressing demand for advanced techniques that can effectively capture the nuanced and multi-faceted nature of emotions within textual data.

Create a deep learning-based text emotion analysis model to accurately detect and classify emotions within textual data. The project encompasses data collection, model selection, optimization, and deployment for real-world applications. Emphasizing precision and versatility, it aims to discern subtle emotional nuances across various contexts with high reliability.

1.3 Existing System

The field of emotion analysis in text has seen significant advancements, categorized into two primary approaches: deterministic models and dimensional models.

1.3.1 Deterministic Models

Deterministic models classify emotions into discrete categories, such as anger, joy, and sadness, often relying on supervised learning techniques. Researchers have utilized machine learning-based methods like Naive Bayes, SVM, and CNNs for emotion classification. Notable datasets, such as GoEmotions (58k Reddit comments annotated with 27 emotions) and Vent (33M social media comments tagged with 705 emotions), have provided a foundation for training and evaluation.

Deep learning-based deterministic models, such as EmoNet and DeepMoji, have further enhanced emotion recognition. EmoNet employs gated recurrent neural networks (GRNNs), while DeepMoji uses emoji-labeled tweets to model affective content. Despite their success, these models struggle with context-dependent and subjective emotional variations.

1.3.2 Limitations of Categorical Emotion Models

Existing systems predominantly employ categorical emotion recognition, which assigns fixed labels (e.g., happy, sad, angry) to textual data. While straightforward, this approach fails to account for:

1. Emotional Ambiguity: Mixed emotional states are not well-represented in discrete categories.
2. Subjectivity: Emotional perception varies significantly across individuals, which is hard to encapsulate with rigid labels.
3. Emotional Ambiguity: Mixed emotional states are not well-represented in discrete categories.

1.3.3 Dimensional Models

Dimensional models represent emotions in continuous spaces, mapping them along axes like Valence (pleasantness) and Arousal (intensity). Russell's Circumplex Model introduced this framework, offering a more nuanced analysis of emotional states. Datasets like EmoBank (annotated on VAD dimensions) and models like BiLSTM-based regressors have demonstrated the utility of dimensional approaches.

Recent research explores transforming emotions from fixed points to probabilistic distributions, addressing subjectivity and fuzzy boundaries. Gaussian Mixture Models (GMMs) have been applied to map emotions as distributions in the Valence-Arousal space, allowing for more personalized emotion analysis.

1.4 Advancements in Deep Learning Frameworks for NLP and Emotion Recognition

Text Classification (TC) involves systematically categorizing textual data into predefined groups based on its intrinsic features and characteristics. By leveraging automated analysis, TC efficiently examines and assigns pre-established categories to text, playing a crucial role in processing and extracting meaningful information from unstructured and raw textual data [1, 2, 3]. TC can be broadly categorized into three system types: rule-based, machine-learning-based, and hybrid approaches [4, 5].

Many researchers have concentrated on enhancing the effectiveness of Natural Language Processing (NLP) tasks by developing advanced Deep Learning (DL) architectures [6, 7]. Notably, methods like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) have achieved significant results, with CNN demonstrating the best average accuracy of 77.4% [8]. These DL advancements have been instrumental in improving the accuracy and efficiency of NLP applications, while ongoing research continues to explore hybrid DL models for further performance optimization.

Combining CNN and BiLSTM within a single framework enables more comprehensive data processing. CNN effectively captures spatial features, whereas BiLSTM focuses on temporal and contextual relationships [9]. Several studies have implemented hybrid CNN-BiLSTM architectures [10, 11], often incorporating word embedding techniques such as Word2Vec, GloVe, and fastText. These models typically utilize an embedding size of 300, Softmax activation, dropout rates of 0.5, recurrent dropout rates of 0.4, the SGD optimizer, 50 training epochs, a filter size of 512, and a kernel size of 3. By integrating embedding layers, CNN max pooling, BiLSTM, and dense layers, these hybrid models have achieved an accuracy rate of 82% [12].

Through rigorous experimentation and comparative evaluation on benchmark datasets, we endeavor to provide valuable insights into the strengths, limitations, and optimal configurations of CNN and BiLSTM-based models for text emotion analysis. By shedding light on the most effective techniques and methodologies, this research aims to contribute toward the advancement of emotion-aware applications and systems, fostering more nuanced and empathetic interactions in the digital realm.

Building on these advancements, this study also explores emerging approaches to emotion recognition, particularly in the context of microblogging platforms like Twitter and Facebook, where informal, user-generated text often presents unique challenges. Microblogs, characterized by their brevity and casual language, contain high-density sentiment-bearing terms that are valuable for understanding collective human behavior [13]. However, such data also introduces complexities, including linguistic variations (e.g., ”thanxxx” for ”thanks”) and contextual ambiguities, where the same text may evoke mixed emotions like joy and annoyance simultaneously.

To address these challenges, we incorporate aspects of dimensional emotion modeling, such as representing emotions in Valence-Arousal space. This approach accounts for fuzzy emotional boundaries and subjective variations, offering a more nuanced understanding of emotions in text. By blending deterministic methods with probabilistic frameworks, we aim to develop intelligent textual emotion recognition systems that are adaptable, scalable, and effective for modern NLP applications.

Chapter 2

Literature Review

2.1 Introduction

Text classification and sentiment analysis are foundational tasks in Natural Language Processing (NLP) that have witnessed substantial improvements through the adoption of deep learning architectures. These advancements have enabled researchers to address complex emotion detection challenges, particularly in multilingual and informal text datasets. By combining state-of-the-art techniques like Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory networks (Bi-LSTMs), and attention mechanisms, recent studies have achieved significant milestones in classification accuracy and robustness. This section reviews key contributions in the domain, focusing on deterministic and hybrid models, as well as emerging probabilistic and dimensional approaches.

2.2 Summary of Relevant Research

Recent research in text emotion analysis spans deterministic and dimensional approaches, focusing on hybrid deep learning architectures and probabilistic models. These studies have significantly advanced the field by improving classification accuracy and addressing the challenges of subjective and ambiguous emotional expressions in textual data.

2.2.1 Deterministic and Hybrid Models

Deterministic and hybrid models have been foundational in advancing text emotion analysis. Deterministic approaches categorize text into fixed emotion labels, while hybrid models combine multiple deep learning techniques to achieve higher accuracy and robustness. These methodologies address challenges in emotion classification by integrating local feature extraction, temporal dependencies, and attention mechanisms, providing a strong baseline for emotion detection tasks.

2.2.1.1 Hybrid Bi-LSTM and Attention Mechanisms:

A study leveraging a Bi-LSTM architecture combined with Word2Vec embeddings, CNNs, and attention mechanisms achieved remarkable performance, with an F1-score of 90.1% on a 13k-instance dataset [12]. By assigning attention weights to significant parts of the text and utilizing Word2Vec embeddings for semantic enrichment, the model demonstrated the advantages of integrating multiple methodologies.

2.2.1.2 Character Embeddings and Multilingual Sentiment Classification:

Salur and Aydin [10] presented a hybrid CNN-BiLSTM model that outperformed standalone models by incorporating character embeddings and pre-trained embeddings like GloVe and FastText. This architecture, tested on Turkish GSM-related tweets, achieved an accuracy of 82.14%, underscoring the value of hybrid deep learning approaches for languages with complex morphological structures.

2.2.1.3 Deep Learning for Low-Resource Languages:

Naqvi et al. [11] addressed sentiment analysis in Urdu, utilizing CNNs for feature extraction and RNNs for capturing temporal dependencies. With a Bi-LSTM ATT model achieving an accuracy of 77.9% and an LSTM model with Samar embeddings reaching 85.16% precision, the study highlighted the potential of tailored embeddings for low-resource languages.

2.2.2 Dimensional Emotion Models

Dimensional models provide a continuous representation of emotions, moving beyond fixed categories to capture emotional nuances in multidimensional spaces. These frameworks, such as Valence-Arousal models, focus on representing the intensity and pleasantness of emotions, offering a more flexible and subjective analysis. Emerging approaches use probabilistic distributions to address fuzzy boundaries and overlapping emotional states.

2.2.2.1 Challenges with Deterministic Models:

Traditional emotion analysis systems classify emotions into fixed categories, which limits their ability to handle subjective variations and fuzzy emotional boundaries. For example, texts like

"The virus is spreading" can evoke fear, sadness, or both, depending on the context and reader's perception.

2.2.3 Context-Aware Models:

Ghafoor [14]. (2023) developed the TERMS model, which employs Gaussian Mixture Models (GMMs) for soft assignment of emotional perceptions in valence-arousal space. Evaluated on 4,000 Twitter messages, TERMS achieved Pearson correlation coefficients of 0.6 for valence and 0.3 for arousal, outperforming prior models like DeepMoji and SemEval-2018 The TERMS (Textual Emotion Recognition in Multidimensional Space) model addresses these challenges by mapping emotions into continuous Valence-Arousal dimensions using Gaussian Mixture Models (GMMs). By representing emotions as probability distributions, TERMS captures overlapping and ambiguous emotional states, making it particularly effective for analyzing informal and opinionated data from social media platforms.

2.3 Comparative Studies

A recent survey by Al Maruf [15]. (2024) consolidated insights into text-based emotion recognition. They noted the predominance of categorical models and identified gaps in dataset diversity and annotation consistency. Their analysis underscored the potential of pooled datasets and dimensional approaches for improved emotion modeling.

2.4 Critical Analysis of Existing Research

The landscape of emotion recognition in text has seen substantial advancements, especially with the integration of deep learning models such as CNN, BiLSTM, and transformers. However, despite these strides, several limitations remain, particularly in terms of the complexities inherent in emotional expression and the adaptability of current models across varied domains.

One prominent limitation is the over-reliance on categorical emotion models, which limit the representation of emotions to predefined labels. This approach is common in systems like Woebot (Gabriels[16]), which categorizes emotional states but fails to capture the continuous and multifaceted nature of human emotions. The categorical model restricts the system's capacity to recognize subtle variations in emotions, which is critical for applications like cognitive behavioral therapy (CBT). On the other hand, some advanced systems, such as the work of Ghafoor [14] have begun incorporating dimensional models, specifically Valence-Arousal-Dominance (VAD) spaces, to better represent the complexity of emotions. These models, which predict emotional states on a continuous scale rather than discrete labels, hold great promise in addressing the nuanced nature of emotional expression.

Furthermore, many existing emotion recognition models suffer from challenges related to dataset biases. For example, Al Maruf [15] highlighted issues of unbalanced datasets and a lack of standardized annotations, which can hinder the performance and generalizability of emotion recognition models. The lack of domain adaptability and inconsistency in dataset structures remain significant challenges, especially when models trained on specific domains are applied to new, unrelated domains.

Additionally, existing models often neglect the integration of contextual information, which is vital for effective emotion recognition, particularly in settings like therapeutic interventions. For instance, systems like Youper and Wysa, although effective, utilize discrete emotional categories without leveraging contextual understanding of emotional progression over time . Context-aware models, like TERMS (Ghafoor [14]), aim to solve this issue by incorporating context-specific classifiers that track emotions across interactions and adapt to the emotional dynamics of a conversation.

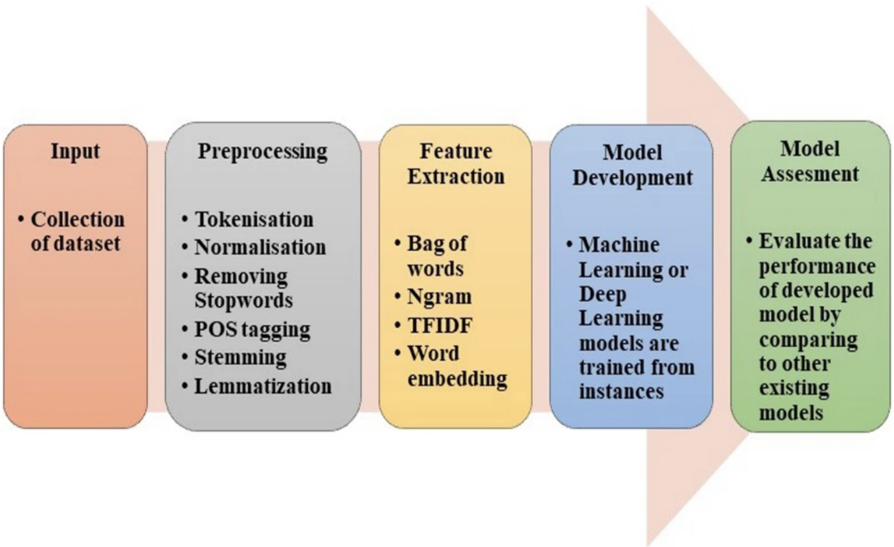
2.5 Connection to the Current Research

This study builds upon the strengths of hybrid models while addressing their limitations by exploring dimensional frameworks. By integrating deterministic CNN-BiLSTM models with the context-aware and probabilistic capabilities. The current research builds upon these findings by introducing a transformer-based model that captures a broader range of emotional dimensions, specifically targeting valence, arousal, and dominance (VAD). Unlike previous categorical systems, this model offers continuous emotion tracking, allowing for finer-grained emotional understanding. By leveraging this transformer-based model, the research seeks to overcome the limitations of fixed emotion categories, providing more precise emotion recognition that is essential for applications like cognitive behavioral therapy (CBT) where emotional subtleties are critical.

Chapter 3

Proposed Work

This research seeks to enhance sentiment analysis in text classification by refining the hybrid CNN-BiLSTM framework and incorporating dimensional emotion modeling to improve the detection of nuanced emotional states. By capitalizing on the complementary strengths of CNN for feature extraction and BiLSTM for capturing sequential dependencies, the study expands the methodology to include probabilistic emotion representations, offering a comprehensive approach to analyzing diverse textual data. Additionally, the research emphasizes developing a deep learning model based on Bidirectional Encoder Representations from Transformers (BERT), optimized for predicting Valence-Arousal-Dominance (VAD) scores in textual inputs. These VAD scores provide a detailed understanding of emotional dimensions within text, supporting applications in sentiment analysis, emotional intelligence, and natural language processing. The proposed model fine-tunes a pre-trained BERT architecture to generate precise VAD scores for various levels of textual granularity, including words, phrases, and sentences.



3.1 Datasets

3.1.1 Twitter Dataset-1

The dataset was created by merging data from DailyDialog, ISEAR, and Emotion-Stimulus, resulting in a well-balanced dataset with five categories: joy, sadness, anger, fear, and neutral. It primarily consists of short text messages and conversational utterances.

- **Test Dataset:** Test Data (GitHub)
- **Train Dataset:** Train Data (GitHub)
- DailyDialog Dataset
- Emotion Stimulus Data (Ottawa)
- ISEAR Dataset (Papers with Code)

3.1.2 Dataset-2

AI Emotion Dataset (HuggingFace)

An example from the dataset is shown below:

”text”: ”im feeling quite sad and sorry for myself but ill snap out of it soon”, ”label”: 0

The dataset includes the following fields:

- **text**: A string representing the input text.
- **label**: A classification label with potential values such as sadness (0), joy (1), love (2), anger (3), fear (4), and surprise (5).

3.1.3 Dataset 3

This dataset consists of four emotions- 0 means Angry,1 means happy,2 means sad,3 means neutral.

- Hindi Emotion Data

3.1.4 Dataset 4

The proposed dataset combines EmoBank, GoEmotions, and ISEAR, resulting in a pool of 75,503 samples with balanced VAD distributions. It integrates categorical-to-dimensional label transformations using the NRC-VAD dictionary and excludes CrowdFlower for independent evaluation, ensuring diverse, domain-spanning data for robust model generalization.

Dataset	Size	Label type
EmoBank	10,000	Native VAD labels
GoEmotions	58,000	Categorical
ISEAR	7,503	Categorical
CrowdFlower	7,500	Categorical

FIGURE 3.1: Dataset 4

3.2 Labeling

- **Categorizing Labels:** The original emotion labels (e.g., `'joy'`, `'fear'`, `'anger'`, `'sadness'`, `'neutral'`) are mapped to unique integer values using a predefined encoding dictionary (e.g., `{'joy': 0, 'fear': 1, ...}`). This numerical representation simplifies the data, making it suitable for computation and feeding into machine learning models.

- **Integer Label Transformation:** The emotion labels in the training and testing datasets (`data_train.Emotion` and `data_test.Emotion`) are replaced with their corresponding integer labels using list comprehension. This step converts the string labels into numerical format, enabling mathematical operations and compatibility with machine learning frameworks.

- **One-Hot Encoding:** After converting the emotion labels to integers, they are further encoded into one-hot vectors. For example:

  - `'joy'` → 0 → `[1, 0, 0, 0, 0]`
  - `'fear'` → 1 → `[0, 1, 0, 0, 0]`

  One-hot encoding treats the labels as categorical variables, ensuring no unintended ordinal relationships between the labels. This representation is crucial for classification tasks where the model should not assume a ranking among the labels.

- **Purpose:** These steps ensure that the emotion labels are properly formatted and structured for machine learning. This allows the neural network to:

  - Interpret the labels correctly.
  - Learn the relationships between input text and emotion categories effectively.
  - Perform accurate emotion classification.

3.3 Preprocessing

Text preprocessing plays a crucial role in text classification by applying various techniques to prepare and refine textual data for effective analysis. It is essential for diverse applications across languages and domains, as raw text data often requires transformation into a machine-readable format for further processing. This step aims to address inconsistencies and eliminate potential issues that might affect subsequent analysis. Since textual data is not always in its ideal form, preprocessing is vital to ensure the labeled data is clean and structured before moving to the next phase. This study incorporates several key preprocessing techniques to improve data quality and enhance its suitability for classification tasks.

3.3.1 Data Cleaning

- **Removing Hashtags and Usernames:** This step eliminates hashtags (e.g., `#example`) and usernames (e.g., `@user`) from the text data, as they typically do not contribute to the emotion analysis task and might add noise to the data.
- **Removing Punctuation:** Punctuation marks like commas, periods, and exclamation points are removed from the text, as they do not typically convey emotion-related information and might interfere with subsequent processing steps.
- **Removing Numbers:** Numerical digits are removed from the text data since they are often irrelevant to emotion analysis tasks and might not contribute meaningfully to the analysis.
- **Lowercasing:** All words are converted to lowercase to ensure consistency in the text data. This step prevents the model from treating the same word with different cases (e.g., `"Hello"` vs. `"hello"`) as different entities.
- **Removing Stopwords:** Commonly occurring words (stopwords) like `"the"`, `"is"`, and `"and"` are removed from the text data as they often do not carry significant meaning in emotion analysis tasks.
- **Stemming:** Words are reduced to their root form (stem) to normalize variations of the same word. For example, `"running"` and `"ran"` are both stemmed to `"run"`, reducing the dimensionality of the feature space and improving generalization.
- **Lemmatization:** Similar to stemming, lemmatization reduces words to their base or dictionary form (lemma). This process ensures that different inflections of the same word are treated as a single entity, further improving the consistency and interpretability of the text data.

3.3.2 Tokenization

- **Tokenization:** Tokenization is the process of converting text into a sequence of tokens or words.
- **Tokenizer Initialization:** A `Tokenizer` object is initialized without specifying any parameters. This object is used to tokenize the text data.
- **Fitting on Texts:** The `fit_on_texts` method of the `Tokenizer` object is called with the entire corpus of texts as input. This method analyzes the texts and creates a vocabulary index based on the frequency of words in the corpus.
- **Texts to Sequences:** The `texts_to_sequences` method is used to convert the text data (`texts_train` and `texts_test`) into sequences of integer indices. Each word in the texts is replaced by its corresponding index in the vocabulary.
- **Word Index:** The `word_index` attribute is used to retrieve a dictionary mapping each word to its corresponding integer index in the vocabulary.
- **Vocabulary Size:** The size of the vocabulary is calculated by adding 1 to the length of the word index. This accounts for the reserved index 0, typically used for padding sequences.

3.3.3 Padding

- **Padding Sequences:** The `pad_sequences` function is used to pad the sequences of integer indices (`sequence_train` and `sequence_test`) to a fixed length (`max_seq_len`).
- **Fixed Input Length:** By specifying `maxlen = max_seq_len`, all sequences are padded or truncated to have the same length. This ensures uniformity in input data size, which is necessary for neural networks.
- **Padding with Zeros:** By default, `pad_sequences` pads sequences with zeros at the beginning (pre-padding) to achieve the desired length.
- **Handling Longest Input:** The maximum sequence length (`max_seq_len`) is set to accommodate the longest input sequence in the dataset, approximately (500,200) words. This ensures no information is lost.

3.3.4 Dimensional Preprocessing for Emotional Features

- **Mapping to VAD Space**: Emotional dimensions — Valence, Arousal, and Dominance were directly calculated for each text sample using the *NRC-VAD dictionary*. Each word or phrase was assigned predefined scores for VAD from the dictionary. For multi-word inputs, the VAD scores were averaged to generate a representative emotional profile for the entire text.

- **Dimensional Feature Augmentation**: The computed VAD scores were concatenated with traditional tokenized text features to create enhanced input vectors. This step allowed the model to incorporate explicit emotional context alongside linguistic features.

3.4 Feature extraction

In Natural Language Processing (NLP), a key challenge lies in building models that can effectively comprehend the hierarchical structure of sentences within textual data. This difficulty is particularly prominent in classification tasks and the extraction of meaningful features. Feature extraction involves identifying and analyzing the distinguishing attributes or characteristics of data. It aims to reduce the dimensionality of output data, thereby simplifying the dataset for more efficient processing. By carefully selecting or combining variables into representative features, the overall data complexity is minimized. In this study, we utilized two widely recognized word embedding techniques—GloVe (Global Vectors for Word Representation) and Word2Vec—for capturing semantic and syntactic relationships in text.

GloVe embeddings are pre-trained word vectors that capture global word co-occurrence statistics in large text corpora. These embeddings encode semantic relationships between words based on their distributional patterns in the corpus. We utilized pre-trained GloVe embeddings, which have been trained on massive text corpora such as Wikipedia and Common Crawl, to obtain dense vector representations for words in our dataset. By leveraging GloVe embeddings, we aimed to capture the semantic similarities and contextual information present in the textual data, thereby enhancing the performance of our text emotion analysis models.

Word2Vec is another popular technique for word embedding, known for its ability to capture semantic relationships between words based on their contextual usage in the text. We employed the Word2Vec algorithm to train word embeddings specific to our dataset, learning vector representations that capture the semantic meaning of words based on their local context within the text. By training Word2Vec embeddings on our dataset, we aimed to capture domain-specific semantics and contextual nuances relevant to text emotion analysis.

3.5 Our Model

The dataset, labeled with sentiment classifications, underwent testing with multiple algorithms, including Neural Networks, LSTM, BiLSTM, and CNN. Before initiating the testing phase, an extensive preprocessing pipeline was implemented to ensure data integrity and relevance. This process involved the removal of elements such as numerical values, hashtags, URLs, emojis, and punctuation to refine the dataset. Following the cleaning phase, feature extraction techniques were applied, which included tokenization and encoding to represent the text effectively.

The data was subsequently partitioned into subsets and processed through padding techniques to standardize sequence lengths, ensuring compatibility with the modeling framework. A comparative analysis was conducted to evaluate the test results against the proposed hybrid model. Initially, a simple BiLSTM model was deployed on the curated datasets to establish baseline performance. The structure and methodology of the BiLSTM model are outlined in the subsequent sections.

Layer	**Output Shape**	**Parameters**
Embedding Layer	(None, 200, 100)	7,530,300
BiLSTM Layer	(None, 256)	234,496
Dense Layer	(None, 6)	1,542

TABLE 3.1: Summary of BiLSTM Model Layers

The construction of the hybrid CNN-BiLSTM model began with the phases of data labeling, preprocessing, and feature extraction, as outlined previously. To represent text data, we utilized an embedding method that converts textual information into low-dimensional numeric vectors, with a padding size fixed at 300. This approach deviates from earlier methodologies to enhance representation and compatibility.

Layer	**Output Shape**	**Parameters**
Embedding Layer	(None, 200, 100)	7,530,300
Convolution1D Layer	(None, 200, 64)	19,264
Pooling Layer	(None, 100, 64)	0
BiLSTM Layer	(None, 256)	197,632
Dense Layer	(None, 6)	1,542

TABLE 3.2: Summary of CNN+BiLSTM(1) Model Layers

And also one more model as follows:

Layer	**Output Shape**	**Parameters**
Embedding Layer	(None, 200, 100)	7,514,600
Convolution1D Layer	(None, 198, 200)	60,200
BiLSTM Layer	(None, 198, 128)	135,680
Dropout	(None, 198, 128)	0
BiLSTM Layer	(None, 128)	98,816
Dense	(None, 50)	6,450
Dense	(None, 50)	2,550
Flatten	(None, 50)	0
Dense	(None, 100)	5,100
Dense	(None, 6)	606

TABLE 3.3: Summary of CNN+BiLSTM(2) Model Layers

The architecture of our model was benchmarked against existing research, summarized in the Table below. Each study customized its architecture to suit the specific characteristics of its dataset. For instance, Jang et al. [12] employed a hybrid CNN-BiLSTM with Word2Vec Skip Gram for analyzing reviews of clothing and camera products. Salur and Aydin [10] utilized a model integrating CNN-BiLSTM with character embeddings and FastText to classify sentiments from tweets by Turkish GSM operator users. Similarly, Soumya and Pramod [17] analyzed sentiment in Malayalam tweets using a hybrid CNN-BiLSTM and CNN-LSTM framework. These models effectively combine CNN layers to extract critical features from the text with BiLSTM layers to grasp contextual relationships between words. The CNN component identifies significant textual patterns using filters and kernels, while BiLSTM processes bidirectional relationships for richer context. To enhance model robustness and prevent overfitting, activation functions like ReLU and dropout mechanisms were implemented [3].

In our model, overfitting was mitigated by incorporating a dropout layer after the initial BiLSTM layer. Dropout randomly deactivates neurons during training, ensuring active neurons prevent the model from overly adapting to the training data. This approach, with a dropout rate set to zero during evaluation, enhances generalization. Furthermore, introducing a dense layer with fewer units in tandem with BiLSTM limits model capacity, reducing the likelihood of overfitting [3].

The combination of CNN and BiLSTM within a unified framework enables a more thorough analysis of the data. While CNN is effective at extracting spatial features, BiLSTM specializes in capturing temporal and contextual information [9]. The hybrid CNN-BiLSTM model begins with data labeling, preprocessing, and feature extraction, as described previously.

To represent text as low-dimensional numerical vectors, the model employs embeddings with padding set at a size of 300. This approach stands out from prior methodologies used by [8, 12, 10, 17], as detailed in the Table below, offering a more nuanced and effective representation [3].

Model	**Embedding**	**F1-Score**	**Accuracy**
CNN BiLSTM[8]	POS Tagging, sentiment vector	81.7%	77.4%
CNN BiLSTM[12]	Word2Vec + Skip Gram[3]	88.0%	87.6%
CNN BiLSTM[10]	Character + FastText	89.0%	82.1%
CNN BiLSTM[17]	Sentiment tagging	75%	85.5%

TABLE 3.4: Comparison of Model Performance with Different Embeddings

3.6 Comparison between CNN-BiLSTM and BERT Models for Emotion Detection

3.6.1 Overview of BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language representation model developed by Google. Unlike traditional models like LSTM, which process data sequentially, BERT processes all words in a sentence simultaneously, using self-attention mechanisms to capture dependencies between words, irrespective of their positions. This bidirectional nature makes BERT particularly effective in understanding context, which is crucial for emotion detection tasks.

3.6.2 BERT for Emotion Detection

In emotion detection, BERT excels due to its ability to understand the context of words and the relationships between them. The typical BERT-based model for emotion detection consists of:

1. **Input Layer**: Tokenized input sequence along with segment and positional embeddings that represent the sentence in context.
2. **BERT Encoder**: The input passes through a stack of transformer layers, where BERT applies bidirectional self-attention to capture the emotional context of the text.
3. **Classification Layer**: A dense layer outputs emotion predictions, classifying the text into different emotional categories based on the features extracted by BERT.

3.6.3 Comparison of CNN-BiLSTM and BERT for Emotion Detection

3.6.3.1 Architecture

- **CNN-BiLSTM**: The CNN-BiLSTM hybrid model combines Convolutional Neural Networks (CNN) for local feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) for sequential context understanding. CNN captures localized emotional cues, while BiLSTM captures the sequential dependencies that may indicate emotional progression or shifts.
- **BERT**: BERT is based on the transformer architecture, utilizing bidirectional self-attention layers. It directly captures the relationships between all words in a sequence, eliminating the need for separate feature extraction layers like CNN. This holistic attention mechanism is highly effective for detecting subtle emotional nuances across the entire text.

3.6.3.2 Pre-training

- **CNN-BiLSTM**: Typically, the CNN-BiLSTM model requires training from scratch or fine-tuning from pre-trained word embeddings like Word2Vec or GloVe. However, these word embeddings may not fully capture the complex emotional context as effectively as BERT's deep contextual understanding.
- **BERT**: BERT is pre-trained on large-scale corpora using tasks like Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), enabling it to learn rich semantic and contextual relationships between words, which is crucial for understanding the context of emotions in text.

3.6.3.3 Context Handling

- **CNN-BiLSTM**: The CNN layer extracts local features such as keywords or phrases indicative of emotional tone. The BiLSTM layer then processes these features sequentially, capturing the progression of emotions over time. However, it processes the input in a fixed order and requires padding for uniform sequence lengths.
- **BERT**: BERT handles context bidirectionally and captures relationships between all tokens in the sequence using self-attention. This makes BERT particularly powerful for emotion detection, as emotions can often span across multiple words, requiring the model to consider the entire context in both directions simultaneously.

3.6.3.4 Performance

- **CNN-BiLSTM**: The hybrid CNN-BiLSTM model is effective in capturing both local features and sequential dependencies. It can achieve good results for emotion detection, especially in simpler tasks. However, it might struggle to capture more complex emotional patterns without extensive fine-tuning.
- **BERT**: BERT typically outperforms CNN-BiLSTM, especially in complex emotion detection tasks. Its ability to capture deep contextual relationships through self-attention leads to superior performance in detecting a wide range of emotions, including nuanced or mixed emotions.

3.6.3.5 Training Time and Resources

- **CNN-BiLSTM**: Training a CNN-BiLSTM model generally requires less computational power and time compared to BERT, making it a more resource-efficient option for simpler tasks or smaller datasets.

- **BERT**: BERT requires substantial computational resources for both pre-training and fine-tuning. Fine-tuning large models like BERT-large can be computationally expensive, but it provides state-of-the-art performance for emotion detection.

3.7 Transformer-Based Approach for Dimensional Emotion Recognition

We employ a deep learning-based approach for dimensional emotion recognition utilizing a transformer-based architecture. Specifically, the model leverages the pre-trained **ALBERT** (A Lite BERT) model , enhanced with a final regression layer designed to predict emotions along the three **VAD** (Valence, Arousal, and Dominance) dimensions. These dimensions are widely recognized for quantifying emotions in terms of their intensity and underlying characteristics. The model undergoes fine-tuning on emotion-annotated datasets, allowing it to effectively map each input sentence to continuous dimensional emotion labels, thereby capturing the nuanced emotional expressions present in text.

3.7.1 Model Architecture and Training

3.7.1.1 ALBERT Model

ALBERT is an optimized version of the BERT architecture that reduces the number of parameters by sharing weights across layers, making it more computationally efficient without compromising performance. The model architecture comprises 12 transformer blocks, each with 768 units, 12 attention heads, and **GELU** (Gaussian Error Linear Units) activations. The ALBERT model is pre-trained on large-scale corpora, including *English Wikipedia* and *Book Corpus*, totaling 25 GB of data, which helps the model capture contextual word relationships and semantic meanings effectively.

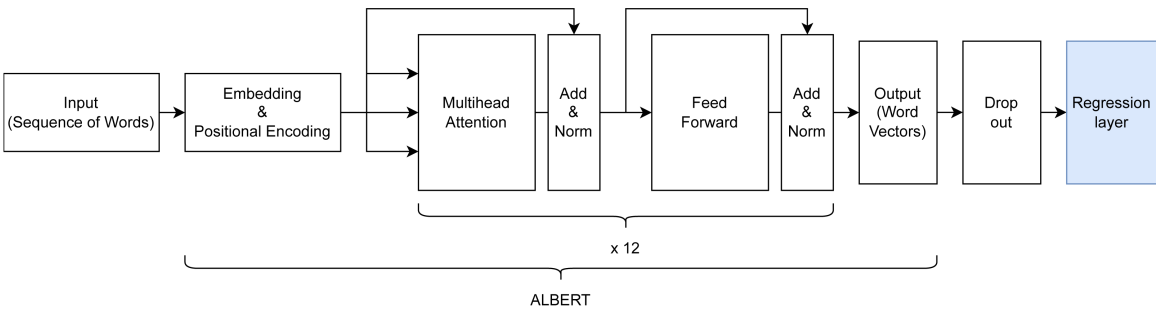


FIGURE 3.2: Albert Model

3.7.1.2 Fine-Tuning for Emotion Recognition

To adapt the ALBERT model for dimensional emotion recognition, a linear regression layer is added on top of the pre-trained encoder. This layer enables the model to predict continuous VAD scores from the textual input, mapping sentences to the respective emotion values. The model was fine-tuned on a dataset split into 78.4% training, 19.6% validation, and 2% test to ensure effective learning and evaluation. This split resulted in over 59,000 training samples and sufficient test samples to avoid overfitting.

3.7.1.3 Training Details

- **Loss Function:** The model uses mean squared error (MSE), a standard loss function for regression tasks, to predict continuous values for the VAD scores.
- **Optimizer:** The AdamW optimizer is employed with an adaptive learning rate of 3e5 to facilitate stable training and prevent overfitting by penalizing large weights.
- **Epochs:** The model is trained over 20 epochs, with a batch size of 32 to ensure proper convergence and avoid overfitting.

3.7.1.4 Final Model Architecture

The final architecture of the model includes:

- **Input Layer:** Accepts input sequences of words (text).
- **Preprocessing Layers:** Handle tokenization, embeddings, and positional encodings for the input text.
- **ALBERT Encoder:** Processes the input through 12 transformer layers to capture contextual relationships and meanings.
- **Dropout Layer:** A dropout rate of 0.1 is applied after the encoder to prevent overfitting.
- **Final Dense Layer:** A dense layer with a linear activation function is used to predict the VAD scores.

Chapter 4

Simulation and Results

4.1 Evaluation of Our Proposed Model

4.1.1 Accuracy with Different Datasets and Embeddings

The performance of different models with various datasets and embeddings is summarized in Table 4.1.

Model	Dataset	Embedding	Accuracy
ML (SVM)	1	Tf-idf vectorizer	72.71%
ML (SVM)	2	Tf-idf vectorizer	87.11%
ML (Logistic Regression)	3	Tf-idf vectorizer	73.4%
BiLSTM	1	Word2Vec (Wikipedia file)	72.68%
BiLSTM	2	GloVe (Twitter)	92.26%
CNN-BiLSTM (1)	1	Word2Vec (Wikipedia file)	73.36%
CNN-BiLSTM (1)	2	GloVe (Twitter)	93.46%
CNN-BiLSTM (2)	1	Word2Vec (Wikipedia file)	73.16%
CNN-BiLSTM (2)	2	GloVe (Twitter)	94%
CNN-BiLSTM (2)	3	Word2Vec	83.6%
BERT	1	WordPiece	82.67%
BERT	2	WordPiece	97.21%
BERT	3	WordPiece	87.21%

TABLE 4.1: Model Performance with Different Datasets and Embeddings

4.1.2 Classification Metrics

The precision, recall, and F1-score for different emotion categories are provided in Table 4.2.

Emotion	Precision	Recall	F1-Score
Anger	0.94	0.95	0.94
Fear	0.94	0.86	0.90
Joy	1.00	0.91	0.95
Love	0.77	1.00	0.87
Sadness	0.98	0.97	0.97
Surprise	0.75	0.99	0.85

TABLE 4.2: Emotion Classification Metrics

4.1.3 Confusion Matrix

Figures 4.1 and 4.2 illustrate the confusion matrices for the CNN-BiLSTM(2) model on Dataset 1 and Dataset 2, respectively.

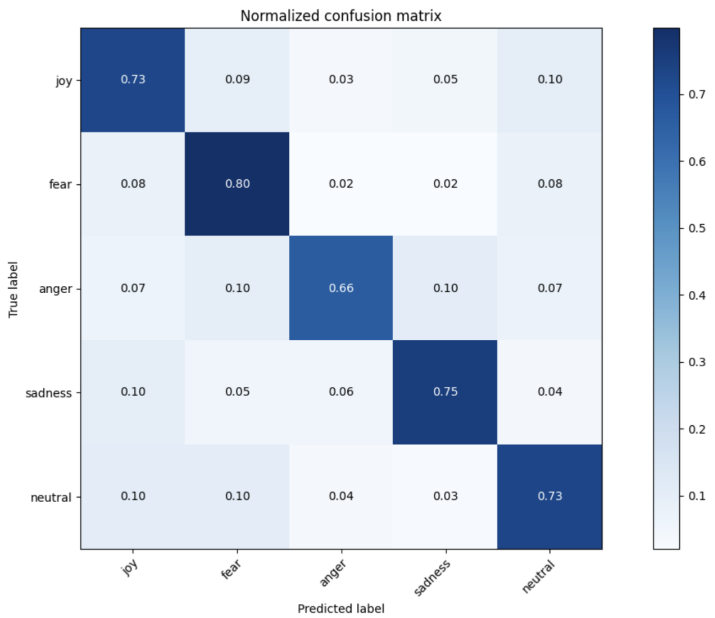


FIGURE 4.1: Confusion Matrix of Dataset 1 for CNN-BiLSTM(2)

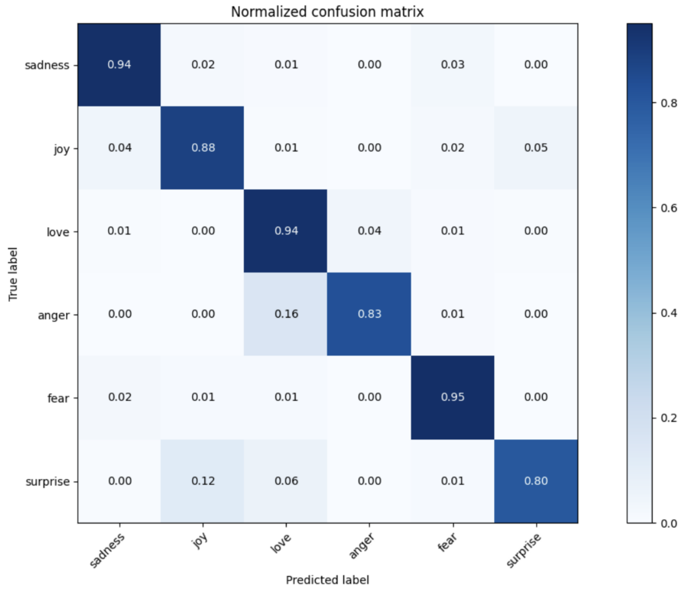


FIGURE 4.2: Confusion Matrix of Dataset 2 for CNN-BiLSTM(2)

4.1.4 Model Training Plots

The training and validation performance of the proposed models during the training process is depicted in the following plots. These plots illustrate key metrics such as accuracy and loss over each epoch, providing insights into the convergence behavior and potential overfitting or underfitting issues.

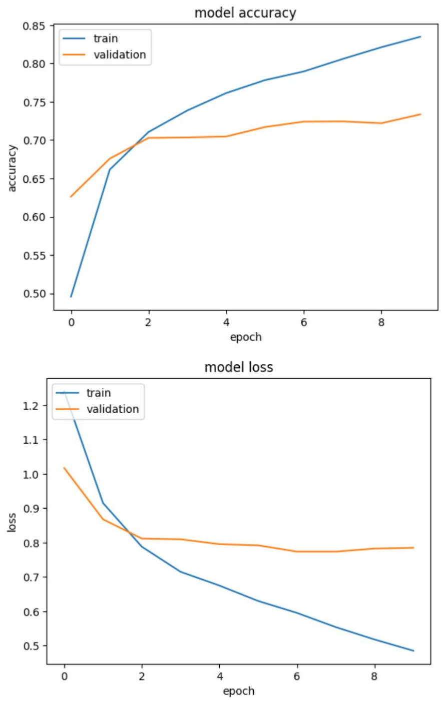


FIGURE 4.3: Training plot of Dataset 1 for CNN-BiLSTM(2)

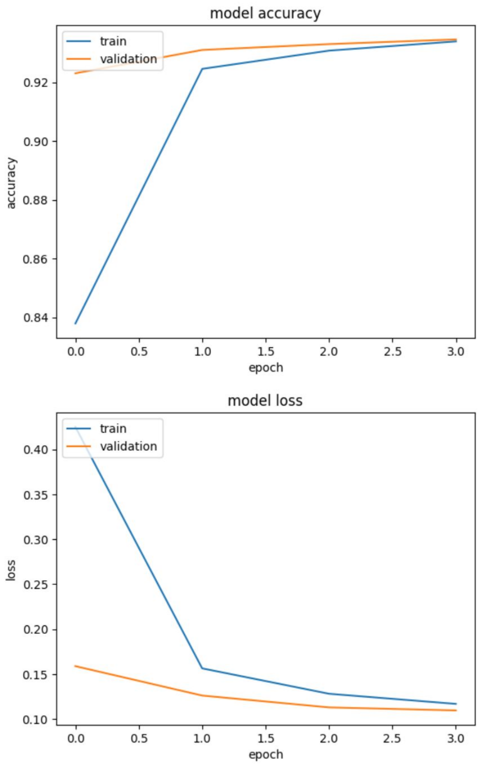


FIGURE 4.4: Training plot of Dataset 2 for CNN-BiLSTM(2)

4.2 Evaluation of Dimensional Emotion Recognition Model

The evaluation demonstrates that the DL-based approach outperforms the rule-based method in emotion recognition tasks. Leveraging metrics such as Mean Squared Error (MSE) and the Pearson correlation coefficient (r), the DL model consistently achieves lower MSE values and higher correlation scores across all datasets. This indicates its superior capability to capture the underlying emotional dimensions. Its advantage lies in training on a combined dataset, allowing it to generalize effectively to unseen data during split evaluations.

	A				**B**			
X	**C**	**D**	**E**	**F**	**C**	**D**	**E**	**F**
P	a	b	c	d	e	f	g	h
Q	i	j	k	l	m	n	o	p
R	q	r	s	t	u	v	w	x
S	y	z	aa	ab	ac	ad	ae	af
T	ag	ah	ai	aj	ak	al	am	an
U	ao	ap	aq	ar	as	at	au	av

TABLE 4.3: Evaluation of Mean Squared Error for Valence, Arousal, Dominance, and Combined VAD Scores in Deep Learning and Rule-Based Emotion Recognition Across Various Datasets.

	A				B			
X	**C**	**D**	**E**	**F**	**C**	**D**	**E**	**F**
P	a	b	c	d	e	f	g	h
Q	i	j	k	l	m	n	o	p
R	q	r	s	t	u	v	w	x
S	y	z	aa	ab	ac	ad	ae	af
T	ag	ah	ai	aj	ak	al	am	an
U	ao	ap	aq	ar	as	at	au	av

TABLE 4.4: Analysis of Pearson Correlation Coefficient for Valence, Arousal, Dominance, and Combined VAD Scores in Deep Learning and Rule-Based Emotion Recognition Across Multiple Datasets.

The comparative evaluation further highlights the effectiveness of the proposed DL-based model for emotion recognition, particularly in the valence-arousal-dominance (VAD) space. Dimensional scores were mapped into five distinct emotion categories using equidistant clusters derived from NRC-VAD values. Precision, recall, and F1 scores on the combined dataset reveal the DL model's significant superiority over the rule-based approach, achieving higher macro, micro, and average F1 scores.

Metric	**DL-Group**			**Rule-Group**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Empty	0.87	0.66	0.75	0.26	0.04	0.07
Threatened	0.87	0.90	0.89	0.31	0.29	0.05
Tranquil	0.79	0.56	0.65	0.19	0.24	0.21
Excited	0.77	0.92	0.84	0.52	0.40	0.07
Rooted	0.89	0.90	0.89	0.61	0.94	0.74
	Macro F1	**Micro F1**	**Average F1**	**Macro F1**	**Micro F1**	**Average F1**
Total	**0.80**	**0.86**	**0.86**	**0.23**	**0.59**	**0.48**

TABLE 4.5: Individual Precision, Recall, and F1 Scores Along with Combined F1 Scores for Transformed Outputs

Chapter 5

Conclusions and Future Work

In terms of NLP observations, the model effectively leverages word embeddings such as GloVe and Word2Vec to capture semantic relationships and contextual nuances within the text data. This aids in accurately interpreting emotions across different contexts and domains. The success of the model showcases the potential of deep learning-based approaches for real-world emotion detection applications, paving the way for more empathetic and sophisticated NLP systems.

We successfully demonstrate the effectiveness of a hybrid CNN-BiLSTM model for text emotion analysis, particularly in the domain of sentiment analysis. The model, combining convolutional and bidirectional long short-term memory networks, exhibits high performance across various datasets. It captures both spatial and temporal features of the text, leading to precise emotion classification with accuracy rates up to 94% using Glove embeddings. The model excels in accurately classifying emotions including joy, sadness, anger, fear , disgust, neutral and surprise.

We have successfully developed a deep learning approach for dimensional emotion recognition using the transformer-based ALBERT model. By fine-tuning the model to predict emotions across the Valence, Arousal, and Dominance (VAD) dimensions, we demonstrated the power of transformers to accurately capture emotions in text. Our model generalizes well across diverse emotional expressions, benefiting from balanced dataset merging. ALBERT's lightweight architecture enables high performance and computational efficiency, making it practical for real-world applications. This work advances emotion recognition by enabling continuous VAD prediction, opening doors for applications in sentiment analysis and emotional health monitoring.

Future work in emotion recognition could explore multimodal approaches, combining text, speech, and facial expressions for more accurate emotion detection. Models could be fine-tuned for specific domains, like healthcare or finance, and adapted for multilingual and cross-cultural contexts. Real-time emotion recognition for interactive systems, personalized emotional profiles, and expanding dimensional emotion recognition (beyond VAD) would further enhance applications. Ethical considerations and model biases should be addressed, while integration with VR/AR could improve user experiences. Additionally, improving accuracy in noisy, informal text and advancing hybrid models with explainability could make emotion recognition systems more reliable and impactful in diverse settings.

Bibliography

[1] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689–42707, 2020.

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in arabic language," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 6629–6643, 2020.

[3] Y. Zhang, T. Wang, and J. Li, "Object classification by effective segmentation of tree canopy using u-net model," *Journal of Advances in Information Technology*, vol. 15, no. 3, pp. 424–432, 2024.

[4] W. Fang, H. Luo, S. Xu, P. E. D. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," *Advanced Engineering Informatics*, vol. 44, p. 101060, 2020.

[5] Z. Liu, C. Lu, H. Huang, S. Lyu, and Z. Tao, "Hierarchical multi-granularity attention-based hybrid neural network for text classification," *IEEE Access*, vol. 8, pp. 149362–149371, 2020.

[6] Q. L. et al., "A survey on text classification: From shallow to deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 1–21, 2020.

[7] F. Zaman, M. Shardlow, S. Hassan, and N. Radi, "Htss: A novel hybrid text summarisation and simplification architecture," *Information Processing and Management*, vol. 57, no. 6, p. 102351, 2020.

[8] K. Pasupa, T. Seneewong, and N. Ayutthaya, "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, pos-tag, and sentic features," *Sustainable Cities and Society*, vol. 50, p. 101615, 2019.

[9] S. Kumar, C. Akhilesh, K. Vijay, and B. Semwal, "A multibranch cnn-bilstm model for human activity recognition using wearable sensor data," *Visual Computing*, vol. 38, no. 12, pp. 4095–4109, 2021.

[10] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[11] U. Naqvi, A. Majid, and S. A. L. I. Abbas, "Utsa: Urdu text sentiment analysis using deep learning methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021.

[12] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, p. 5841, 2020.

[13] A. Bermingham and A. Smeaton, "Classifying sentiment in microblogs: Is brevity an advantage?," in *Proceedings of CIKM*, pp. 1833–1836, 2010.

[14] Y. Ghafoor, J. Jinping, F. H. Calderon, Y.-H. Huang, K.-T. Chen, and Y.-S. Chen, "Terms: Textual emotion recognition in multidimensional space," *Applied Intelligence*, vol. 53, no. 3, pp. 2673–2693, 2023.

[15] A. A. Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, F. Mridha, and Z. Aung, "Challenges and opportunities of text-based emotion detection: a survey," *IEEE Access*, vol. 12, pp. 18416–18450, 2024.

[16] K. Gabriels, "Response to uncertainty in emotion recognition," *Journal of Information, Communication and Ethics in Society*, vol. 17, no. 3, pp. 295–298, 2019.

[17] S. Soumya and K. V. Pramod, "Hybrid deep learning approach for sentiment classification of malayalam tweets," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 891–899, 2022.