Sentiment Extraction from Patient Feedback and Health Records using Deep Learning

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF
MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

ANKIT SINGH

2K23/CSE/30

Under the supervision of

Prof. Rahul Katarya

COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Ankit Singh**, Roll No – **2K23/CSE/30**, student of M.Tech (**Computer Science and Engineering**), hereby declare that the project dissertation titled "**Sentiment Extraction from Patient Feedback and Health Records using Deep Learning**", which is submitted by me to the **Department of Computer Science and Engineering**, Delhi Technological University, Delhi, in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, or other similar title or recognition.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Sentiment Extraction from Patient Feedback and Health Records using Deep Learning**" which is submitted by **Ankit Singh**, Roll No – **2K23/CSE/30**,**Computer Science and Engineering** ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **Prof. Rahul Katarya**

Date: May,2025 SUPERVISOR

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express our sincerest gratitude to **Prof. Rahul Katarya** for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve our targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi — Ankit Singh

Date: May,2025

ABSTRACT

The rapid growth of patient-generated drug reviews has introduced a valuable resource for understanding real-world treatment outcomes and user experiences. This thesis explores deep learning approaches for extracting sentiment from such unstructured feedback, focusing on two architectures: Bidirectional Long Short-Term Memory (Bi-LSTM) and a hybrid Convolutional Neural Network with Bi-LSTM (CNN-BiLSTM).

Unlike many existing studies that depend on static pre-trained embeddings, both models are trained from scratch using task-specific word representations. This enhances their adaptability to domain-specific terminology and nuanced sentiment expressions commonly found in medical reviews.

The experiments are conducted on the publicly available Drug Review Dataset from the UCI Machine Learning Repository, hosted on Kaggle. Reviews are preprocessed and labeled as positive, negative, or neutral based on user-provided ratings. The models are evaluated using standard metrics including accuracy, precision, recall, F1-score, and Cohen's Kappa.

The results indicate that the CNN-BiLSTM model outperforms the standalone Bi-LSTM across all metrics. Its combined ability to extract local features and capture bidirectional context leads to more robust sentiment prediction. This research demonstrates the feasibility of building end-to-end sentiment classifiers tailored to healthcare narratives and lays the groundwork for future extensions such as multilingual analysis, aspect-based sentiment classification, and interpretable AI models.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	**Full Form**
AI	Artificial Intelligence
API	Application Programming Interface
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
F1-score	Harmonic Mean of Precision and Recall
FN	False Negative
FP	False Positive
GloVe	Global Vectors for Word Representation
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
TF	TensorFlow
TN	True Negative
TP	True Positive

Table 1: List of abbreviations used in this thesis.

Chapter 1

INTRODUCTION

In recent years, the proliferation of user-generated healthcare content—ranging from online drug reviews to patient forums and informal clinical narratives—has opened new avenues for understanding patient perspectives on treatments. These unstructured textual sources offer valuable insights into drug efficacy, side effects, and overall patient satisfaction. However, extracting structured meaning from such content requires sophisticated techniques capable of interpreting natural language. Sentiment analysis, which aims to determine the emotional tone (positive, neutral, or negative) of textual data, serves as a powerful tool in this context [1, 2].

Conventional sentiment analysis techniques, including lexicon-based methods, handcrafted rules, and classical machine learning algorithms, often perform well in general-purpose applications [2]. Nevertheless, their effectiveness diminishes when applied to healthcare narratives due to the presence of specialized terminology and nuanced expressions common in medical reviews.

Recent advancements in deep learning have significantly enhanced natural language understanding. Architectures like Long Short-Term Memory (LSTM) networks [3] and Convolutional Neural Networks (CNNs) [4] have demonstrated strong capabilities in learning context and semantic patterns from raw textual input. Bidirectional LSTM (Bi-LSTM) models further extend this by incorporating context from both directions of a sequence. Hybrid frameworks that combine CNNs with Bi-LSTMs have shown promising results in capturing both local phrase-level patterns and long-term dependencies [5, 6].

This thesis presents a deep learning-based approach for sentiment classification of patient-submitted drug reviews. Instead of depending on general-purpose, static word embeddings such as GloVe [7], the proposed models are trained from scratch using healthcare-specific data, enabling them to learn relevant linguistic features directly from the corpus.

This research utilizes the publicly available Drug Review Dataset sourced from the UCI Machine Learning Repository and hosted on Kaggle [8]. This dataset contains patient-written feedback and associated numeric ratings, which are used as proxy sentiment labels. Model performance is assessed using established classification metrics such as accuracy, precision, recall, F1-score, and Cohen's Kappa

All models are implemented using the TensorFlow deep learning framework and trained on Google Colab, leveraging its scalability and ease of experimentation [9, 10].

1.1 Motivation

As patient-centered healthcare continues to gain importance, analyzing the subjective experiences shared by patients has emerged as a critical tool for improving treatment outcomes and tailoring healthcare services. Online platforms such as Drugs.com provide a wealth of real-world feedback on medications, capturing valuable insights related to effectiveness, side effects, and overall satisfaction—often beyond the scope of controlled clinical trials.

Yet, this feedback is inherently unstructured and linguistically diverse, making it difficult to analyze manually at scale. The complexity increases due to mixed sentiments within single reviews, domain-specific terminology, and variability in expression.

To address this challenge, there is a growing need for automated sentiment analysis systems that can understand healthcare-specific language and derive actionable insights from raw textual reviews. This research aims to fill that gap by developing deep learning models that are trained end-to-end on domain-relevant data, without relying on static, pre-trained word embeddings. By learning directly from patient-written content, the models are better equipped to adapt to the nuances of medical narratives and provide more accurate and meaningful sentiment classification. This approach holds promise for supporting healthcare professionals, pharmaceutical companies, and policymakers in interpreting large volumes of patient feedback efficiently and effectively.

1.2 Objectives

The principal aims of this study are:

- To develop and implement deep learning models—namely Bi-LSTM and a CNN-enhanced Bi-LSTM—for sentiment analysis of patient drug reviews.
- To perform training exclusively on patient reviews and associated ratings, without using external static word embeddings.
- For assessing such models on balanced datasets and regular evaluation metrics, such as F1-score, recall,Cohen's kappa, accuracy and precision.
- To conduct comparative analysis between models and investigate the benefits of hybrid architectures over standalone recurrent models.
- To explore potential improvements for sentiment detection in the context of medical text analytics.

1.3 Challenges

Several challenges arise when applying deep learning techniques to sentiment analysis in the healthcare domain:

- **Specialized vocabulary:** Medical text often includes domain-specific terminology and abbreviations that may not be well-represented in general language models.

- **Ambiguous sentiment:** Patient reviews frequently contain mixed or conflicting sentiments (e.g., "The medication worked but gave me headaches"), making classification nontrivial.
- **Imbalanced labels:** Positive reviews tend to outnumber negative or neutral ones, which can lead to biased predictions if not addressed.
- **Limited labeled data:** Annotated medical sentiment datasets are scarce; hence, indirect labeling based on numerical ratings is used.
- **Interpretability:** Deep learning models have traditionally been referred to as black boxes, which limits their acceptability in high-stakes fields like healthcare.

This thesis aims to tackle these issues using models that are both data-driven and tailored to the healthcare domain, emphasizing end-to-end training and minimal reliance on external resources.

1.4 Thesis Organization

- **Chapter 1: Introduction** — Introduces the research problem, motivation, objectives, challenges, provides a high-level overview of the study.
- **Chapter 2: Literature Review** — Reports on current sentiment analysis methods,from conventional methods to deep learning architectures in current times, with emphasis on healthcare applications.
- **Chapter 3: Research Methodology** — Describes the dataset used (Drug Review Dataset from Kaggle), preprocessing pipeline, model architectures, and training strategy.
- **Chapter 4: Results and Discussion** — Presents experimental setup, evaluation metrics, model performance comparison, and qualitative analysis.
- **Chapter 5: Conclusion and Future Scope** — Condenses the main findings,constraints, and proposes future developments like multilingual and aspect-based sentiment analysis.

Chapter 2

LITERATURE REVIEW

2.1 Literature Survey

Over the past two decades, sentiment analysis has grown from rule-based methods and lexicon-driven systems into a robust field heavily influenced by deep learning advancements. Foundational research by Pang and Lee [1] laid the groundwork for supervised sentiment classification using machine learning algorithms. Building upon this, Liu [2] provided an extensive review of techniques based on linguistic rules, statistical methods, and manual feature design.

In medical and health-related contexts, textual data presents unique challenges. Patient narratives tend to include domain-specific terminology and informal language that can hinder traditional models' performance. Colón-Ruiz and Segura-Bedmar [11] demonstrated the value of deep learning architectures—specifically recurrent neural networks—in analyzing Spanish-language tweets related to health. Their work showed that end-to-end neural models are particularly effective for noisy and colloquial text found in real-world health data.

Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber [3], remain central to sequential data modeling. These were later extended into Bidirectional LSTM (Bi-LSTM) networks, which process input in both temporal directions, enhancing context comprehension. In a separate line of work, Kim [12] illustrated that Convolutional Neural Networks (CNNs), originally designed for visual data, could also effectively extract semantic features from text sequences.

Recent efforts have focused on hybrid architectures that blend CNNs with Bi-LSTM models to capture both spatial and temporal patterns in textual data. Zhou et al. [5] further refined this idea by integrating attention mechanisms, allowing the model to emphasize the most salient parts of the input. Gräßer et al. [6] applied such methods to fine-grained sentiment classification in pharmaceutical reviews, showing promising improvements in aspect-specific opinion detection.

Together, these developments form the backbone of modern sentiment analysis and provide a solid foundation for this thesis, which employs deep learning models for classifying sentiment in patient-generated drug reviews.

2.2 Dataset

This study uses a single large-scale dataset: the *Drug Review Dataset*, which is publicly available on Kaggle and originally sourced from the UCI Machine Learning Repository [13]. The dataset comprises over 160,000 user-submitted reviews of various medications, each associated with a numerical rating on a scale of 1 to 10. Along with the review text, each record includes the drug name, the medical condition it was used for, and a "useful count" indicating how many users found the review helpful.

For the purpose of supervised sentiment classification, the numeric ratings are mapped to sentiment labels using a three-class scheme. Ratings from 1 to 4 are categorized as *Negative*, ratings 5 to 6 are labeled as *Neutral*, and ratings from 7 to 10 are marked as *Positive*. This rating-to-label conversion allows the model to learn from implicit sentiment cues, even in the absence of explicit annotations.

The dataset is well-suited for this task due to its substantial volume, diversity of medical conditions, and the use of natural language written by real users. These characteristics make it possible to train deep learning models that can capture domain-specific linguistic patterns and generalize across different contexts in patient feedback.

index	condition	review ▼	rating
18	Bipolar Disorde	"This medication helped me sleep, but eventually it became ineffective as a sleep aid. It also helps me calm down when in severe stress, anxiety, or panic."	6
12	Weight Loss	"This is a waste of money. Did not curb my appetite nor did it make me feel full."	1
17	Acne	"So I was on Ginanvi for about 3 months before I switched over to this pill due to the high cost of Ginavi (I don't have insurance). Ginanvi had cleared up my acne in the short time I used and I was hoping this birth control would not mess that up. I was wrong. After 2-3 months of using this pill my skin went crazy with horrible breakouts. Now, 9 months later I still have horrible acne. If you want a pill that helps with acne this is NOT it. It also made my period "weird". I would have little to no period one month and a heavy one the next . The month that I had a heavier flow my boobs would hurt a lot. I also noticed that at the beginning of taking it , I would be moody and extremely irritable all the time . It's time for me to switch."	3
13	Schizoaffective Disorde	"No problems, watch what you eat."	10
20	Insomnia	"If I could give it a 0, I would absolutely do so. Started at 50mg, and felt WIRED. Wanted to get up and clean the house! Bumped it to 100mg, less wired, but still wide awake all night. Bumped to 150, with the same lack of effect. MD informed me after this dose it becomes less effective for sleep, so why even bother. 15 years of trying different sleep medications and alternatives, and this, I can say for sure, was the LEAST effective I have ever come across. At it's low price point, feel free to give it a try, and maybe you will be luckier? Everyone's sleep conditions are different. But if you get hyper after benadryl, expect the same reaction to this drug."	1
16	Smoking Cessation	"I smoked for 50+ years. Took it for one week and that was it. I didn't think it was possible for me to quit. It has been 6 years now. Great product."	10
11	Vaginal Yeast Infection	"Honestly its day one on the 3 day treatment. Yes it burns a bit and it does leak out if you dont lay down after insertion. But im faithful it will work."	6
10	Hot Flashes	"Holy Hell is exactly how I feel. I had been taking Brisdelle for 1.5 years. The hot flashes did indeed subside - however, the side affects of this medicine coupled with the fact Noven was acquired by YET another pharmaceutical company - YOU CAN'T PLACE A REP IN THE AREA, DISTRIBUTE YOUR DRUGS, AND THEN FIRE HER-AND NOT REPLACE THEREFORE there is NO medicine or support here. You dumped this drug in the Dr's hands and walked away. After calling Sebula - you act like you don't even care. You have made it impossible to obtain this. I happen to think this is illegal. I just decided to wean myself off this and Premarin. It has been nothing short of a nightmare. If you don't need this drug- DON'T START. Seriously."	1
14	Insomnia	"Ditto on rebound sleepless when discontinued. I have done very strange things with no memory including taking additional Ambien. It has helped me sleep when under extreme stress but watch out. Now I am trying to learn how to sleep naturally."	2
19	Urinary Tract Infection	"After just 1 dose of this ciprofloxacn, I felt 99% better."	10
15	Narcolepsy	"A doctor in the ER prescribed me 200 mg of Provigil when I was first diagnosed with Narcolepsy. It didn't seem to have any effect on me at all. Then I went to see my sleep doctor and he prescribed me 250 mg of Nuvigil. It finally allowed me to stay awake for an entire day. But if I went out somewhere I would be able to come home and take a nap and be fully awake again. Taking that with Xyrem makes me feel 98%..not quite 100, but still pretty great."	9

Figure 2.1: Sample entries from the Drug Review Dataset showing review text, condition, and rating fields.

2.3 Problem Statement

Although patient drug reviews offer a wealth of information, analyzing them for sentiment poses several difficulties. Traditional approaches, including lexicon-based techniques and classifiers built on static word embeddings, often fail to capture the complex, domain-specific expressions common in healthcare text. For example, phrases like "no improvement" may be misclassified if models lack contextual understanding.

Another layer of complexity arises from the presence of mixed sentiments within individual reviews. A patient might describe both the benefits and side effects of a medication, making it challenging to assign a single sentiment label. This ambiguity is less common in domains like product reviews, where sentiment is usually more uniform.

The central research challenge addressed in this thesis is the design of deep learning models that can extract sentiment from free-form patient narratives without using generic, pre-trained embeddings. Instead, the models are trained directly on the healthcare datasets, allowing them to learn specialized patterns and vocabulary that better reflect the medical context. The goal is to create end-to-end systems capable of nuanced and accurate sentiment classification based solely on the data itself.

Chapter 3

PROPOSED METHODOLOGY

3.1 Research Design

This research adopts an empirical methodology to develop and evaluate deep learning models for sentiment classification of patient drug reviews. The focus is on two architectures: a Bi-LSTM network and a hybrid model that integrates a Convolutional Neural Network (CNN) with Bi-LSTM. Both models are trained from scratch using domain-specific data, without relying on external word embeddings.

The overall research process is structured into the following five stages:

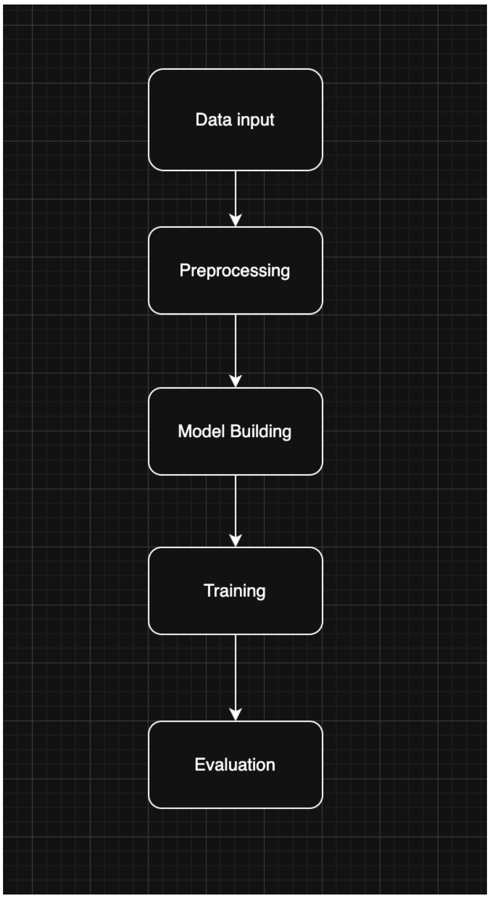

Figure 3.1: Flowchart illustrating the end-to-end research design process.

1. **Data Acquisition** — Collecting and consolidating review data from the Kaggle-hosted Drug Review Dataset.

2. **Data Preprocessing and Feature Engineering** — Cleaning and preparing the review text for input into neural networks through steps such as tokenization, padding, and sentiment label assignment.

3. **Model Design** — Constructing the deep learning architectures with layers tailored to capture both local and sequential text features.

4. **Model Training** — Using TensorFlow and Keras within the Google Colab environment to train the models end-to-end with GPU support.

5. **Evaluation and Analysis** — Measuring model performance with standard classification metrics and conducting error analysis.

This structured approach ensures reproducibility, facilitates iterative experimentation, and supports rigorous evaluation of model effectiveness in extracting sentiment from unstructured medical feedback.

3.2 Dataset Description

This study uses a single large-scale dataset of patient-submitted drug reviews, publicly available on Kaggle [13]. The dataset, originally sourced from the UCI Machine Learning Repository, comprises over 160,000 records split across training and testing files. Each record includes several fields: the name of the medication, the medical condition for which it was used, the textual review, a numeric rating (1–10), the date of submission, and a count of how many users found the review helpful.

For the purpose of sentiment classification, the numeric ratings are converted into categorical labels. A three-class mapping scheme is applied:

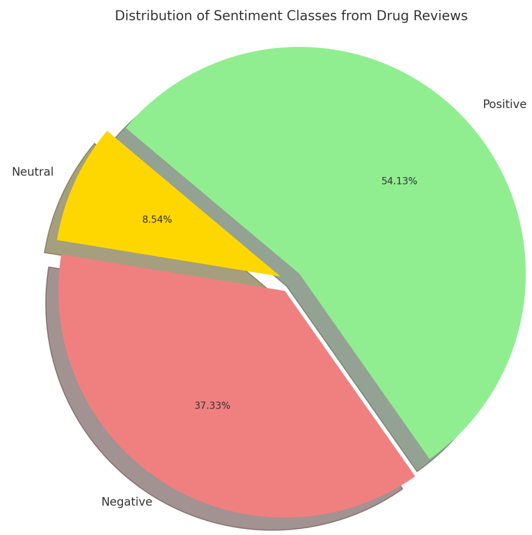

Figure 3.2: Distribution of sentiment classes (Positive, Neutral, Negative) based on user ratings.

- 1 to 4 are labeled as **Negative**.
- 5 to 6 are classified as **Neutral**.

- 7 to 10 are marked as **Positive**.

This approach enables weak supervision, where sentiment labels are inferred from numerical ratings rather than manually annotated. The diversity of medical conditions and the real-world language used by patients in the reviews make this dataset a strong foundation for training deep learning models aimed at extracting sentiment.

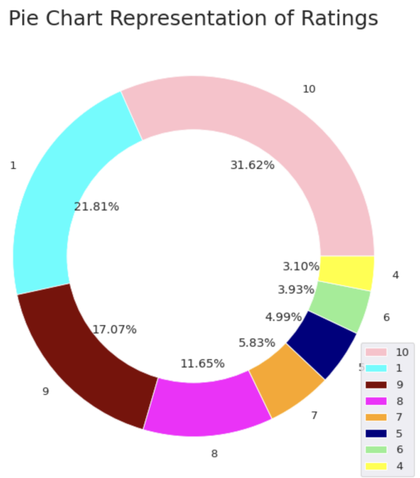

Figure 3.3: Distribution of sentiment classes derived from patient review ratings.

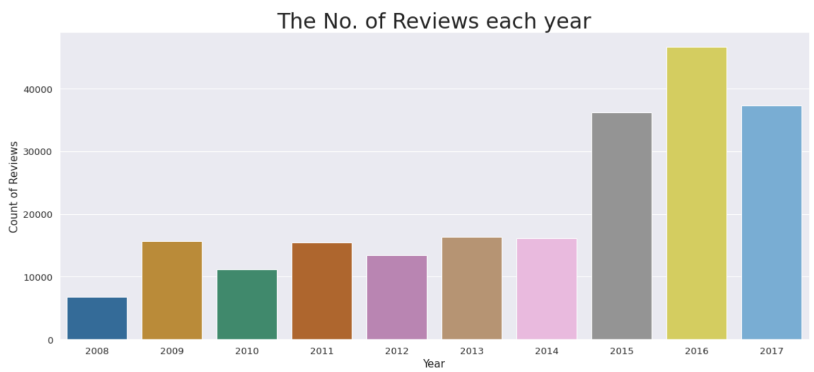

Figure 3.4: Bar chart showing the number of user-submitted drug reviews per year from 2008 to 2017.

3.3 Data Preprocessing

Preprocessing plays a critical role in converting unstructured textual reviews into a format that deep learning models can efficiently interpret and learn from. The raw patient feedback undergoes the following sequential transformations:

- **Text Normalization (Lowercasing):** All characters are converted to lowercase to reduce redundancy and maintain uniformity in word representation.

- **Stopword Removal:** Frequently occurring words that contribute minimal semantic value—such as "is", "the", and "an"—are removed to focus on more meaningful terms.
- **Punctuation and Special Character Removal:** All symbols, punctuation, and non-alphanumeric characters are stripped from the text using regular expressions.
- **Tokenization:** Each review is split into individual tokens (words), enabling word-level analysis and embedding.
- **Lemmatization:** Words are reduced to their root forms (e.g., "treating" becomes "treat"), helping to decrease vocabulary size while preserving core meanings.
- **Sequence Padding:** Since input reviews vary in length, sequences are padded with zeros to ensure consistent input dimensions for the neural networks.

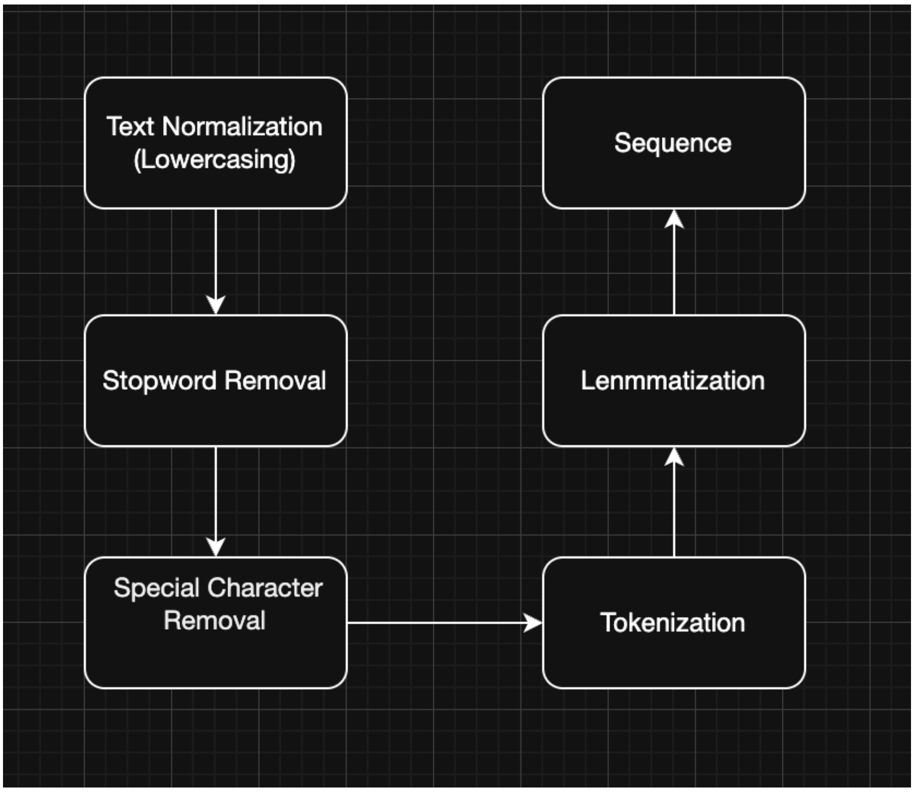

Figure 3.5: Preprocessing pipeline applied to the DrugReview dataset.

After preprocessing, the cleaned reviews are transformed into sequences of integers using a Keras tokenizer. This tokenizer constructs a vocabulary index based on the corpus and encodes each word in the text accordingly. The resulting numerical sequences are then passed into the embedding layer as model input.

3.4 Model Architectures and Working Principles

3.4.1 Convolutional Neural Networks (CNN)

CNNs are highly effective for extracting spatial features from data. In sentiment analysis, CNNs apply filters over text embeddings to detect local n-gram patterns that may correlate with sentiment. This makes them particularly useful for identifying key phrases or expressions.

The architecture starts with the embedding layer that converts tokens into dense vectors. The 1D convolutional layer is next, scanning embeddings with ReLU activation. Max-pooling is applied for reducing dimensions while preserving strong features. Final classification is performed with the dense layer with softmax activation. Categorical cross-entropy loss is used to train the model.

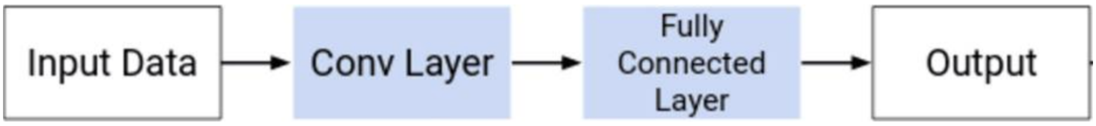

Figure 3.6: CNN Architecture for Sentiment Classification.

3.4.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. These networks are especially suited for tasks like sentiment analysis, where understanding the context over time is crucial. LSTMs manage information flow using a set of internal gates, each of which plays a distinct role in updating and preserving memory [3, 14].

Each LSTM cell includes the following key mechanisms:

- **Forget Gate:** This gate controls which parts of the previous cell state should be discarded. It takes the current input x_t and the previous hidden state h_{t-1} as input and computes the forget vector:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3.1}$$

- **Input Gate:** This gate determines what new information will be added to the cell state. It has two components:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{3.2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3.3}$$

- **Cell State Update:** The updated cell state C_t is computed by combining the previous state and the new candidate values:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{3.4}$$

- **Output Gate:** This gate generates the output hidden state h_t based on the updated cell state:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{3.5}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{3.6}$$

In these equations, σ represents the sigmoid activation function and tanh is the hyperbolic tangent function. These activations ensure that the LSTM selectively retains and updates information over time. For multi-class sentiment classification tasks, the LSTM output is typically passed to a dense layer with softmax activation. The model is optimized using categorical cross-entropy as the loss function.

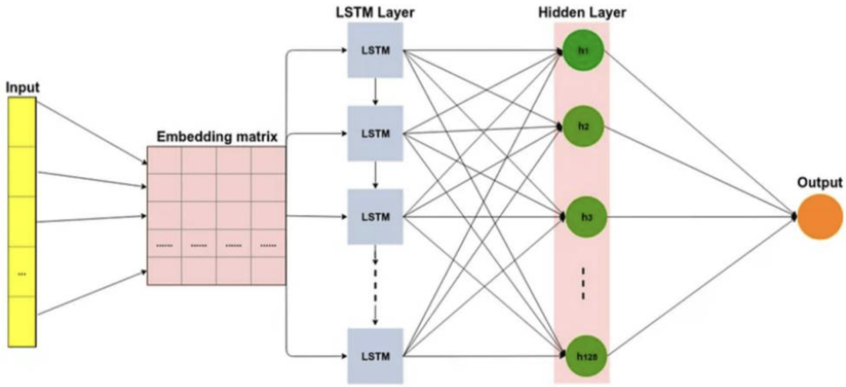

Figure 3.7: LSTM Cell Architecture.

3.4.3 Bidirectional LSTM (Bi-LSTM)

Bi-LSTM networks add to the standard LSTM architecture two parallel layers: one processes the sequence in one direction, and another in the other direction. The two perspectives enable the model to leverage past and future context information, which is especially beneficial in sentiment analysis tasks where surrounding context greatly influences meaning [15].

Let p_t^f and p_t^b represent the output vectors from the forward and backward LSTM layers at time t, respectively. The final output p_t is computed by combining these two vectors:

$$p_t = p_t^f + p_t^b \tag{3.7}$$

where:

- p_t is the resulting probability distribution over sentiment classes,
- p_t^f is the forward LSTM's output at time t,
- p_t^b is the backward LSTM's output at time t.

The final concatenated output is passed to a dense layer followed by softmax activation to perform multi-class sentiment classification. This architecture has been proven effective in numerous NLP tasks, particularly when the full sentence context is important [16, 17, 18].

3.4.4 Bidirectional LSTM Model

This architecture is built to capture both forward and backward dependencies in text sequences. The layers include:

- **Embedding Layer**: Transforms input tokens into trainable vector embeddings.
- **SpatialDropout1D Layer**: Regularizes the embedding layer to prevent overfitting.
- **Bidirectional LSTM Layer**: Employs 64 memory units to capture context in both directions.

- **Dense Output Layer**: Applies softmax to classify sentiment into three classes.

This model is particularly effective in capturing full context when interpreting patient-written reviews [3].

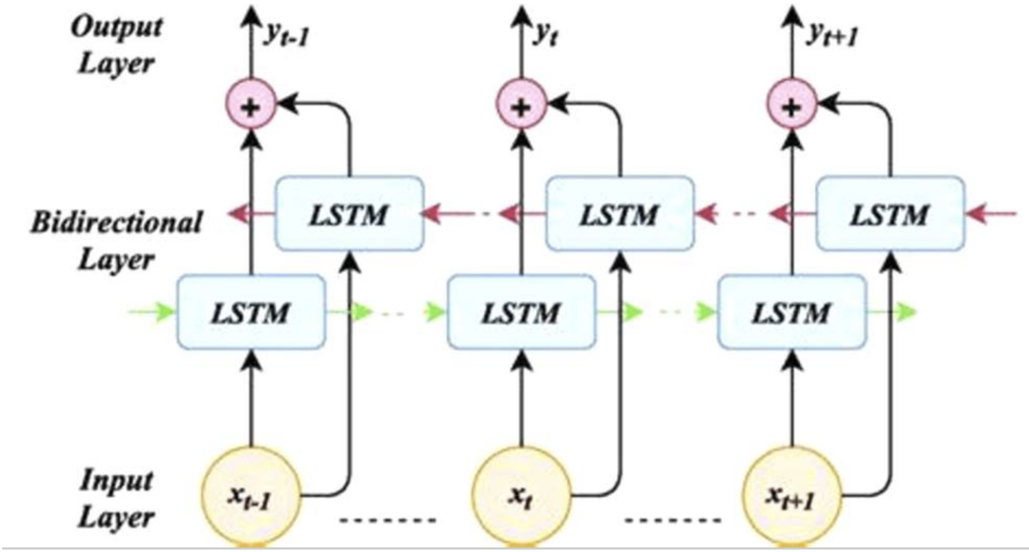

Figure 3.8: Bi-LSTM Architecture: Combines forward and backward LSTM layers for enhanced context understanding.

3.4.5 Hybrid CNN-BiLSTM Architecture

The CNN-BiLSTM model combines the local feature extraction of CNN with the contextual modeling of Bi-LSTM. Initially, convolution and pooling layers capture prominent local patterns such as key phrases. The Bi-LSTM layer then interprets these patterns within the full sequence context.

This hybrid approach benefits from CNN's ability to extract spatial features and Bi-LSTM's strength in modeling sequential dependencies. The model uses ReLU in the convolution layer, softmax in the output layer, and is trained using categorical cross-entropy.

3.4.6 CNN-BiLSTM Hybrid Model

To improve performance, a hybrid model is designed by incorporating a CNN before the Bi-LSTM. The architecture includes:

- **Embedding Layer**: Converts input tokens into embeddings.
- **1D Convolutional Layer**: Extracts local textual features using filters.
- **MaxPooling Layer**: Reduces spatial dimensions and highlights salient features.
- **Bidirectional LSTM Layer**: Learns dependencies in both forward and backward directions.
- **Dense Output Layer**: Performs multi-class classification using softmax.

This model leverages the strengths of both CNN and Bi-LSTM for robust sentiment classification [12, 5].

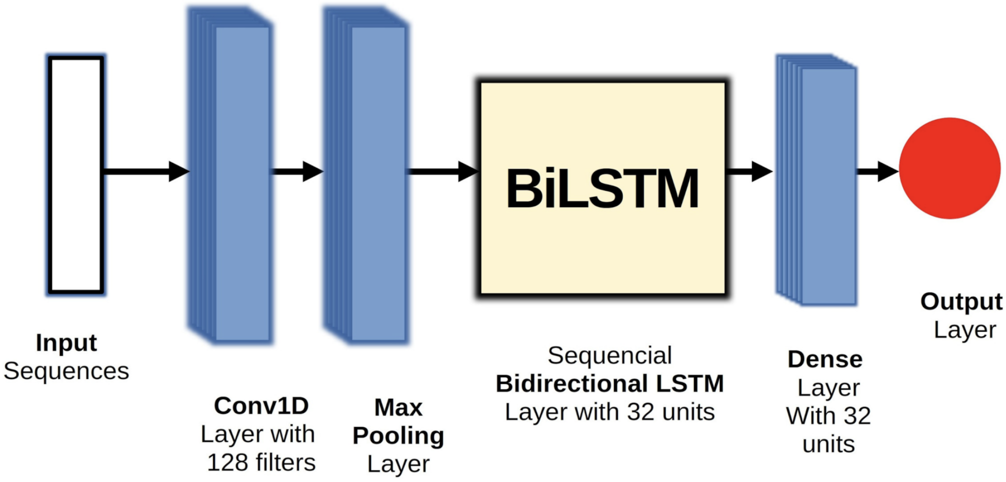

Figure 3.9: Hybrid CNN-BiLSTM Architecture

3.5 Training and Evaluation Strategy

The dataset is partitioned into training and testing sets using a 75:25 ratio. Additionally, a portion of the training data is further split to create a validation set, which is used to track the model's performance during training and prevent overfitting.

Training is conducted using the Adam optimization algorithm, paired with the categorical cross-entropy loss function—well-suited for multi-class classification problems. Each model is trained for 10 to 20 epochs, with a batch size of 64. Key hyperparameters such as the learning rate and dropout rate are selected through empirical experimentation to maintain a trade-off between convergence and overfitting control.

To assess the effectiveness of the models, several evaluation metrics are employed:

1. **Accuracy**: The ratio of correctly predicted instances to the total number of predictions.

2. **Precision**: The proportion of true positive predictions among all predicted positives.

3. **Recall**: The ability of the model to capture all relevant positive cases.

4. **F1-Score**: The harmonic mean of precision and recall, providing a balance between them.

5. **Cohen's Kappa**: A statistical measure that evaluates inter-rater agreement, adjusted for chance agreement.

These metrics are calculated on the test dataset to assess generalization performance.

Chapter 4

RESULTS AND DISCUSSION

4.1 Experimental Setup

All experimental evaluations were carried out on Google Colab [19], utilizing its built-in GPU acceleration capabilities to expedite training. The implementation of the models was done using TensorFlow and Keras [20]. The dataset was partitioned with 75% used for training and the remaining 25% for testing the model's generalization performance.

Both the Bi-LSTM and the CNN-BiLSTM models were trained on sequences derived from patient reviews that had been tokenized and padded to a uniform length. The word embedding layer was initialized randomly and trained so that the model learned the word representations from the medical context automatically. All the models were trained for more than 15 epochs with a batch size of 64 on the Adam optimizer and loss function categorical cross-entropy since it is suitable for multiclass classification.

4.2 Model Performance

Model performance was evaluated using a set of standard classification metrics: precision, recall, accuracy, F1-score, and Cohen's Kappa.These metrics provide a comprehensive assessment of how well each model captures sentiment across multiple classes. The classification task involves three sentiment categories—positive, neutral, and negative—and thus benefits from metrics that balance performance across imbalanced class distributions.

Table 4.1 provides a side-by-side comparison of the Bi-LSTM and CNN-BiLSTM models. The CNN-BiLSTM architecture consistently outperforms the standalone Bi-LSTM across all metrics. This improvement highlights the advantage of combining convolutional layers—which effectively extract local text patterns—with bidirectional LSTM layers that capture long-range dependencies in the review text.

Table 4.1: Performance Comparison of Bi-LSTM and CNN-BiLSTM Models

Metric	**Bi-LSTM**	**CNN-BiLSTM**
Accuracy	0.875	**0.942**
Precision	0.861	**0.938**
Recall	0.858	**0.944**
F1-Score	0.859	**0.941**
Cohen's Kappa	0.823	**0.912**

To further understand the training dynamics of the Bi-LSTM model, Figure 4.1 shows the training and validation loss and accuracy curves. The plots reveal a consistent downward trend in loss and a steady improvement in accuracy, indicating stable learning behavior and minimal overfitting.

4.2.1 Bi-LSTM Training Analysis

The performance of the Bidirectional Long Short-Term Memory (Bi-LSTM) model is illustrated using its training and validation curves in Figure 4.1. The model was trained for five epochs using padded review sequences, and the loss and accuracy values were monitored throughout.

As shown in the plots, both training and validation losses steadily decline, while accuracy increases over each epoch. This trend reflects stable convergence, indicating that the model is successfully learning semantic patterns from the data without overfitting.

The Bi-LSTM architecture is preferred since it can handle input sequences both in forward and backward directions so that it can capture contextual dependencies better than unidirectional models. This two-directional context processing is especially valuable in natural language applications like sentiment analysis, when the meaning of a word or phrase will more than likely be based in prior and subsequent words.

Despite its overall strong performance, the model demonstrates some class-level confusion, particularly between neutral and negative sentiments, as later shown in the confusion matrix. Nevertheless, the Bi-LSTM remains a robust baseline for deep learning in sentiment classification.

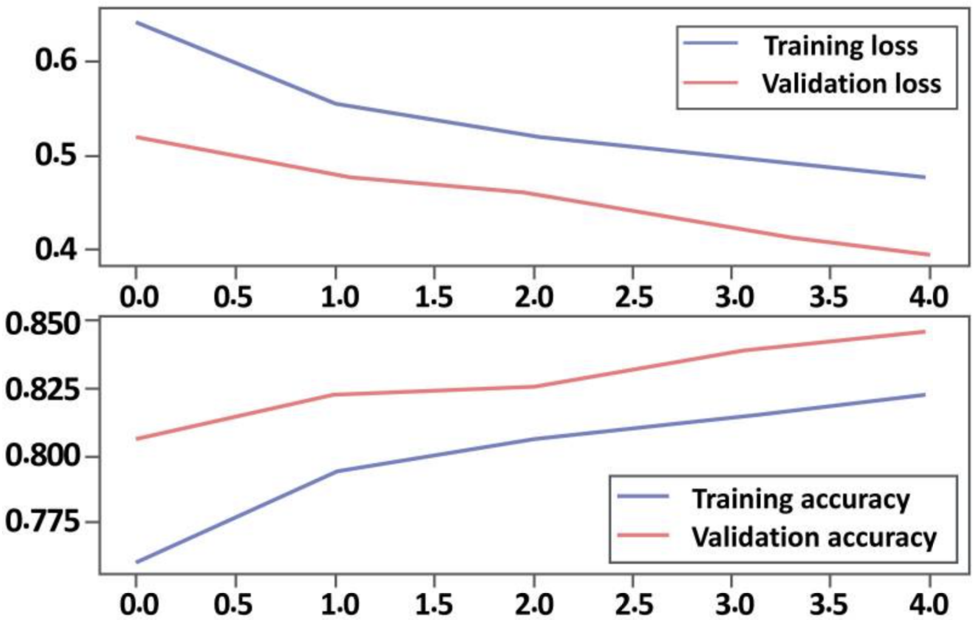

Figure 4.1: Training and validation loss and accuracy curves for the Bi-LSTM model.

In addition, the confusion matrix in Figure 4.2 provides insight into the class-wise performance of the Bi-LSTM model. It reveals a strong prediction capability for the positive class, while showing moderate confusion between neutral and negative classes. This is expected in sentiment tasks where user reviews often contain mixed tones.

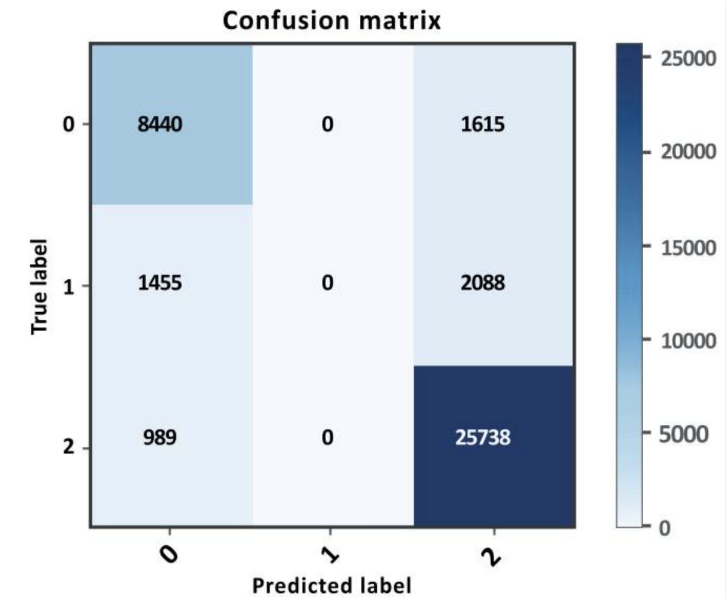

Figure 4.2: Confusion matrix for the Bi-LSTM model.

4.2.2 CNN-BiLSTM Training Analysis

To evaluate the CNN-BiLSTM model, training and validation performance over epochs is plotted in Figure 4.3. The graphs show a steady decline in both training and validation loss, accompanied by a consistent improvement in accuracy. This indicates that the model is able to learn effectively while generalizing well to unseen data.

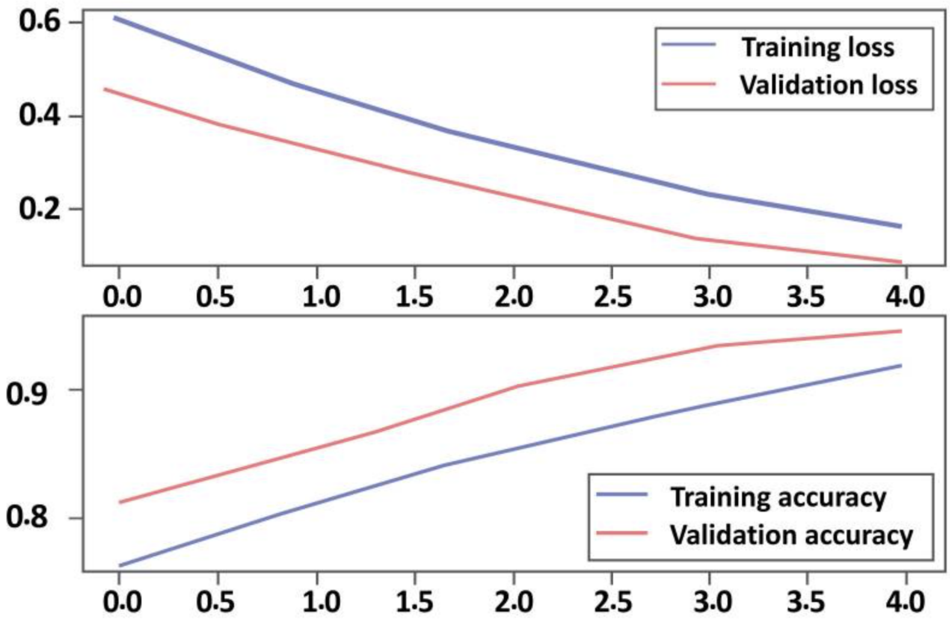

Figure 4.3: Training and validation loss and accuracy curves for the CNN-BiLSTM model.

Figure 4.4 presents the confusion matrix for the CNN-BiLSTM model. The model shows high precision in identifying positive and neutral sentiment classes, with relatively minimal confusion. Compared to the Bi-LSTM-only configuration, this hybrid model reduces misclassification, especially in the neutral class, which is typically more ambiguous.

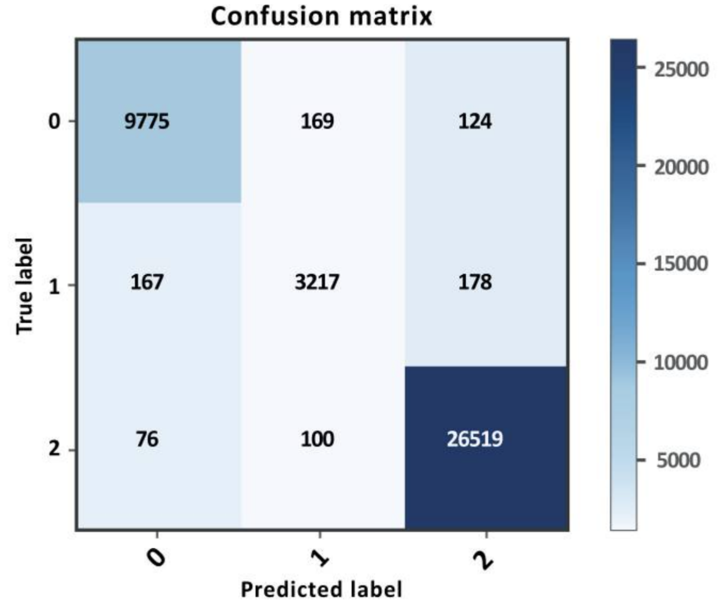

Figure 4.4: Confusion matrix for the CNN-BiLSTM model.

Together, these evaluations demonstrate that while both models perform well, the CNN-BiLSTM architecture is better suited for capturing nuanced sentiment patterns in real-world patient reviews.

4.3 Comparative Analysis

As indicated in Table 4.1, the CNN-BiLSTM model consistently achieved superior scores across all evaluation metrics when compared to the standalone Bi-LSTM. The convolutional layer in the hybrid model plays a key role in identifying important local features, such as specific medical expressions and keyword clusters. These features are then fed into the Bi-LSTM layer, which excels at capturing the contextual dependencies across the review text.

Cohen's Kappa values further reinforce this observation by demonstrating stronger inter-rater agreement between the CNN-BiLSTM model's predictions and the actual labels. The improvements across all performance indicators suggest that integrating both local and global text analysis contributes meaningfully to sentiment classification tasks in the medical domain.

4.4 Discussion of Findings

The results demonstrate that the combination of CNN and Bi-LSTM enhances the model's ability to interpret reviews that contain mixed sentiments. For example, many patients express satisfaction with treatment outcomes but also highlight adverse side effects. The CNN-BiLSTM model is better equipped to handle such nuances due to its capacity for both local pattern recognition and contextual sequence analysis.

Another important insight is the effectiveness of training the embedding layer from scratch using domain-specific data. Despite not using external pre-trained embeddings such as GloVe or Word2Vec, the models achieved high accuracy, illustrating that large and representative healthcare datasets can support robust model training.

Figures 4.5 and 4.6 illustrate the training and validation trends over epochs, showing stable and improving model behavior without overfitting.

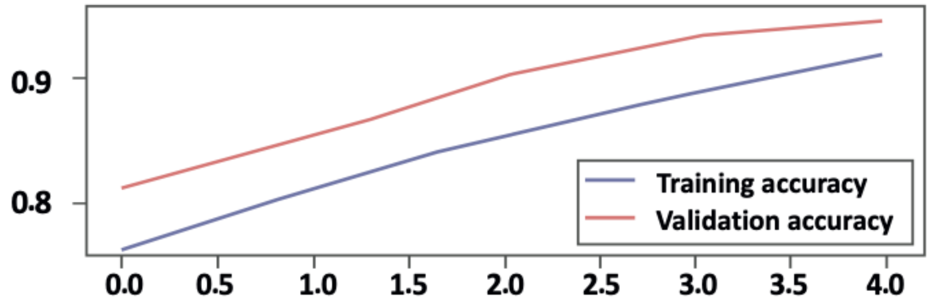

Figure 4.5: Training and validation accuracy over epochs.

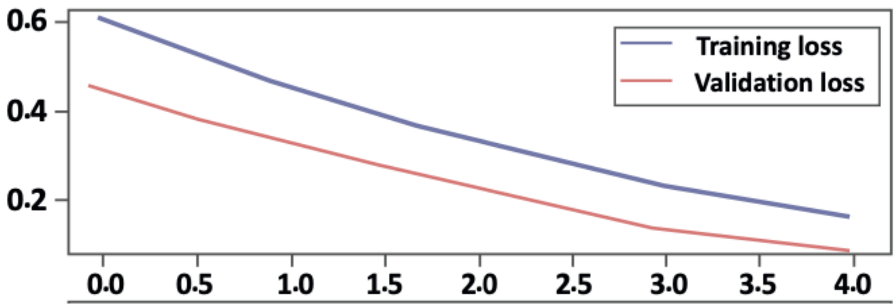

Figure 4.6: Training and validation loss over epochs.

Furthermore, the confusion matrix for the CNN-BiLSTM model, shown in Figure 4.7, reveals balanced performance across sentiment categories, with minimal misclassification.

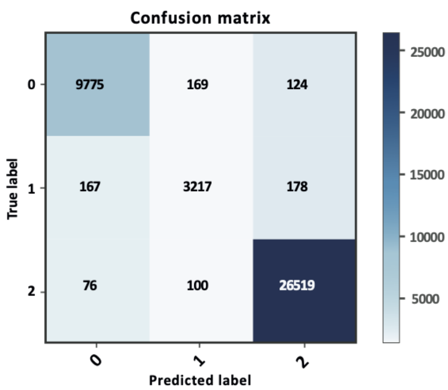

Figure 4.7: Confusion matrix for CNN-BiLSTM model.

Overall, the CNN-BiLSTM architecture presents a scalable and domain-sensitive solution for extracting sentiment from patient narratives. Its ability to generalize across diverse review patterns makes it suitable for integration into real-world clinical feedback systems.

Chapter 5

CONCLUSION FUTURE WORK and SOCIAL IMPACT

5.1 Summary of Findings

This thesis investigated the use of deep learning architectures for sentiment classification in patient-generated drug reviews, with an emphasis on learning from scratch without reliance on pre-trained embeddings. Two models were designed and tested: a standalone Bi-LSTM network and a hybrid CNN-BiLSTM model that integrates convolutional and sequential processing layers.

Empirical results showed that the CNN-BiLSTM architecture consistently outperformed the Bi-LSTM model across all key evaluation metrics—accuracy, precision, recall, F1-score, and Cohen's kappa. The hybrid design benefited from the strengths of both CNNs, which identify localized patterns, and Bi-LSTMs, which capture long-range dependencies in text sequences. This allowed the model to more effectively interpret complex sentiment expressions, especially in reviews containing both praise and criticism.

Significantly, the models learned word embeddings directly from the dataset, eliminating the need for external linguistic resources such as GloVe or Word2Vec. This not only simplified the training process but also allowed the embeddings to adapt specifically to the healthcare domain, improving overall model robustness and relevance.

5.2 Limitations

Although the models demonstrated strong performance, there are certain constraints that may affect their broader applicability:

- **Language and Cultural Bias:** The dataset predominantly consists of English-language reviews submitted by users from a limited demographic. As a result, the findings may not generalize well across different languages or cultural expressions of sentiment.

- **Class Distribution Bias:** While the dataset was balanced artificially for the purpose of training and evaluation, natural sentiment distributions in user feedback are often skewed—especially toward positive sentiment. This could affect the model's effectiveness when deployed in real-world applications.

- **Lack of Interpretability:** Deep neural networks often function as black-box models, making it challenging to interpret their decision-making process—particularly concerning in high-stakes fields like healthcare.

- **Sentiment Resolution:** The current models classify sentiment on a coarse-grained scale (positive, neutral, negative), but do not distinguish between different aspects of sentiment such as emotional distress, drug effectiveness, or side effect severity.

5.3 Future Work

Several avenues remain open for future research, each offering the potential to improve the precision, scope, and interpretability of sentiment analysis systems in the healthcare domain:

- **Aspect-Based Sentiment Analysis (ABSA):** Future models could be designed to detect sentiment at the aspect level—for example, classifying sentiment related to effectiveness separately from side effects or affordability.
- **Model Explainability:** Incorporating mechanisms such as attention layers, or external interpretability tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), would improve transparency and increase the trustworthiness of the models.
- **Domain-Specific Pretraining:** Although this study avoided the use of pre-trained embeddings, building domain-specific word vectors from large-scale medical literature could provide a middle ground—retaining contextual depth without losing domain relevance.
- **Multilingual Analysis:** Expanding the methodology to handle feedback in multiple languages could make the system applicable to more diverse patient populations worldwide.
- **Use of Transformer-Based Models:** Models based on transformer architectures such as BERT or BioBERT, fine-tuned on medical reviews, could be evaluated against the current LSTM-based frameworks to assess potential gains in performance and generalization.

In conclusion, this work demonstrates the value of training deep learning models directly on domain-specific data for sentiment extraction. The success of the CNN-BiLSTM model provides a strong foundation for building intelligent tools capable of analyzing large volumes of patient feedback. Such systems have the potential to support clinicians, researchers, and policy-makers by converting unstructured reviews into actionable insights.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] B. Liu, *Sentiment analysis and opinion mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012, vol. 5, no. 1.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1746–1751.

[5] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. ACL*, 2016, pp. 207–212.

[6] A. Gräßer, P. Kallumadi, C. Malberg, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying attention mechanisms," in *Proc. IEEE BHI*, Chicago, IL, USA, 2018, pp. 50–53.

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1532–1543.

[8] Kaggle, "Drug reviews dataset," [Accessed: May 20, 2025], 2025, https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon.

[9] TensorFlow, "An end-to-end open source machine learning platform," [Accessed: May 20, 2025], 2025, https://www.tensorflow.org/.

[10] Google Colab, "Colaboratory: A google research project," [Accessed: May 20, 2025], 2025, https://colab.research.google.com/.

[11] M. Colón-Ruíz and I. Segura-Bedmar, "Deep learning for sentiment analysis of spanish health tweets," *Journal of Biomedical Informatics*, vol. 100, p. 103324, 2020.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014, pp. 1746–1751.

[13] J. Li, "Drug review dataset," Kaggle, 2018, https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[15] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[16] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm networks,” in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, vol. 4. IEEE, 2005, pp. 2047–2052.

[17] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” *Proceedings of NAACL*, 2016.

[18] Y. Wang, M. Huang, L. Zhao, and X. Zhu, “Attention-based lstm for aspect-level sentiment classification,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 606–615.

[19] Google, “Colaboratory - google research,” Online, 2022, https://colab.research.google.com/.

[20] Google Brain Team, “Tensorflow: An end-to-end open source machine learning platform,” Online, 2022, https://www.tensorflow.org.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: Sentiment Extraction from Patient Feedback and Health Records using Deep Learning

Total Pages: 29 **Name of Scholar:** Ankit Singh

Supervisor: Prof. Rahul Katarya

Department: Computer Science and Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: Turnitin **Similarity Index:** 14% **Total word count: 6,225**

Date: ______________________

Candidate's Signature **Signature of Supervisor**

check plag report.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98324632

Submission Date

May 29, 2025, 1:29 PM GMT+5:30

Download Date

May 29, 2025, 1:32 PM GMT+5:30

File Name

check plag report.pdf

File Size

6.3 MB

29 Pages

6,225 Words

37,742 Characters

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)
- Crossref database
- Crossref posted content database

Match Groups

57 Not Cited or Quoted 14%
Matches with neither in-text citation nor quotation marks

5 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources

4% Publications

12% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **57** Not Cited or Quoted 14%
  Matches with neither in-text citation nor quotation marks
- **5** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
4% Publications
12% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	VIT University on 2025-04-09	<1%
2	Internet	www.mdpi.com	<1%
3	Internet	www.dspace.dtu.ac.in:8080	<1%
4	Submitted works	National College of Ireland on 2024-08-12	<1%
5	Internet	link.springer.com	<1%
6	Internet	arxiv.org	<1%
7	Internet	dspace.dtu.ac.in:8080	<1%
8	Submitted works	University of Central Lancashire on 2024-05-06	<1%
9	Internet	www.dline.info	<1%
10	Internet	iemtronics.org	<1%

11 Internet

dokumen.pub <1%

12 Internet

www.coursehero.com <1%

13 Submitted works

Brunel University on 2025-04-16 <1%

14 Submitted works

Obudai Egyetem on 2025-05-14 <1%

15 Submitted works

Liverpool John Moores University on 2024-07-07 <1%

16 Submitted works

Nanyang Technological University on 2024-03-03 <1%

17 Submitted works

University of Portsmouth on 2025-05-16 <1%

18 Submitted works

University of Edinburgh on 2023-04-14 <1%

19 Submitted works

Vilnius Gediminas Technical University on 2022-11-04 <1%

20 Internet

bora.uib.no <1%

21 Internet

pmc.ncbi.nlm.nih.gov <1%

22 Submitted works

Associatie K.U.Leuven on 2025-05-21 <1%

23 Publication

Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%

24 Submitted works

University of Bristol on 2023-05-04 <1%

25 Internet

dspace.uiu.ac.bd <1%

26 Internet

www.jetir.org <1%

27 Publication

Leon, Javier. "Forecasting Imported Fruit Prices in the United States Using Neural... <1%

28 Submitted works

University of Hertfordshire on 2023-08-25 <1%

29 Internet

infodemiology.jmir.org <1%

30 Internet

oa.upm.es <1%

31 Submitted works

Liverpool John Moores University on 2025-01-05 <1%

32 Submitted works

The University of Manchester on 2024-04-19 <1%

33 Submitted works

University of Essex on 2024-08-27 <1%

34 Submitted works

University of London External System on 2022-04-02 <1%

35 Submitted works

University of Ulster on 2025-05-06 <1%

36 Internet

ethesis.nitrkl.ac.in <1%

37 Internet

ijeponline.org <1%

38 Submitted works

Al-Nahrain University on 2022-12-04 <1%

39 Submitted works

Asia Pacific University College of Technology and Innovation (UCTI) on 2025-02-21 <1%

40 Submitted works

Birkbeck College on 2025-05-16 <1%

41 Submitted works

National Institute of Technology, Hamirpur on 2019-05-30 <1%

42 Submitted works

University of Hertfordshire on 2025-01-05 <1%

43 Submitted works

University of Stirling on 2025-04-09 <1%

44 Submitted works

University of Sydney on 2025-05-23 <1%

45 Submitted works

Wright State University on 2023-07-21 <1%

46 Internet

bright-journal.org <1%

47 Submitted works

cit on 2025-05-23 <1%

48 Submitted works

dsscwellington on 2025-02-08 <1%

49 Internet

ijeecs.iaescore.com <1%