

Emotion Detection from Medical Conversations: Employing Deep Learning Models with Text and Audio Features

submitted by,
Rahul Lugun
Class Roll No: 002210503011
Exam Roll: MCA244011
Master of Computer Application

under the supervision of
Prof. (Dr.) Sivaji Bandyopadhyay,
Dr. Dipankar Das

in the
Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University

2024

Faculty of Engineering and Technology Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “**Emotion Detection from Medical Conversations: Employing Deep Learning Models with Text and Audio Features**” has been carried out by Rahul Lugun (Reg. No.: 163603 of 2022-2023 , Roll No.: 002210503011, Exam Roll: MCA244011), under my guidance and supervision and be accepted in partial fulfilment for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024

Prof. (Dr.) Sivaji Bandyopadhyay
Project Supervisor
Dept. of Computer Science &
Engineering
Jadavpur University

Dr. Dipankar Das
Project Supervisor
Dept. of Computer Science &
Engineering
Jadavpur University

Prof. (Dr.) Nandani Mukhopadhyay
Head of the Department
Dept. of Computer Science &
Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council of
Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology

Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project entitled “**Emotion Detection from Medical Conversations: Employing Deep Learning Models with Text and Audio Features**” is a bonafide record of work carried out by Rahul Lugun (Reg. No.: 163603 of 2022-2023 , Roll No.: 002210503011, Exam Roll: MCA244011), in fulfilment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessary endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date:

Signature of Examiner 2

Date:

Faculty Of Engineering and Technology

Jadavpur University

DECLARATION OF COMPLIANCE WITH ACADEMIC ETHICS

I hereby declare that this project entitled “Emotion Detection from Medical Conversations: Employing Deep Learning Models with Text and Audio Features” is an outcome of my genuine and independent work, as part of my Master of Computer Application program

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced

Rahul Lugun

Registration No.: 163603 of 2022-2023

Examination Roll No.: MCA244011

Thesis Title: - Emotion Detection from Medical Conversations: Employing Deep Learning Models with Text and Audio Features

Signature with Date

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank **Dr. Dipankar Das, Department of Computer Science & Engineering** and **Prof. (Dr.) Sivaji Bandyopadhyay, Department of Computer Science & Engineering**, my supervisors, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honoured that I for this opportunity work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support

Last, but not the least, I would like to thank my family and friends for staying by my side when I needed them.

Rahul Lugun
Roll. No.: 002210503011
Examination Roll No.: MCA244011
Department of Computer Science & Engineering
Jadavpur University

LIST OF FIGURES

Figure	Name	Page No.
1	LSTM Network architecture	7
2	BiLSTM architecture	8
3	BERT architecture	9
4	confusion matrix LSTM (Audio)	16
5	confusion matrix LSTM (Text)	17
6	confusion matrix BiLSTM (Audio)	18
7	confusion matrix BiLSTM (Text)	19
8	confusion matrix BERT (Text)	20

LIST OF TABLES

Table	Name	Page No.
1	Distribution by Disease Category	10
2	Disease Category by Sentiment	10
3	Snapshot of the dataset	12
4	Sentiment Analysis result	15
5	LSTM prediction error	17
6	BiLSTM prediction error	19
7	BERT prediction error	20

TABLE OF CONTENT

List of Figures.....	i
List of Tables.....	i
Abstract.....	ii
Chapter 1: Introductions.....	1
1.1.Overview.....	1
1.2.Background & Motivation.....	1
1.3.Objective.....	1
1.4.Contribution.....	2
1.5.Outline of the report.....	2
Chapter 2: Literary Survey.....	3
Chapter 3: Methodology.....	4
3.1.Problem Statement.....	4
3.2.Proposed Technique / Approach.....	4
3.2.1. Audio Processing.....	5
3.2.2. Text Processing.....	6
3.2.3. Models: LSTM, BiLSTM, BERT.....	7
Chapter 4: Experiments & Results.....	10
4.1.Dataset.....	10
4.1.1. Dataset Statistics.....	10
4.2.Experimental Setup.....	12
4.2.1. Audio.....	12
4.2.2. Text.....	12
4.3.Results.....	15
4.3.1. Sentiment Analysis.....	15
4.4.Error analysis.....	16
4.4.1. LSTM.....	16
4.4.2. BiLSTM.....	18
4.4.3. BERT.....	20
4.5.Discussions.....	21
Chapter 5: Conclusion and Future work.....	23
References.....	24

ABSTRACT

This study explores sentiment analysis techniques applied to both audio and textual data in the context of medical consultations. The sentiment analysis pipeline incorporates shallow level sentiment analysis using NLTK's VADER, followed by deep learning models for both audio and text data. For audio sentiment analysis, Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models are employed, leveraging audio features extracted from the recordings. In contrast, textual sentiment analysis utilizes LSTM, BiLSTM, and BERT models, which are trained on preprocessed transcriptions of the audio content. The effectiveness of these models is assessed through comprehensive experimentation and comparative analysis. The findings shed light on the performance and suitability of various sentiment analysis approaches for medical consultation data, contributing to the advancement of sentiment analysis techniques in healthcare settings.

CHAPTER 1: INTRODUCTION

1.1 Overview

In recent years, the digitization of healthcare data has led to an explosion of available information, including audio recordings of medical conversations between patients and healthcare providers. These conversations contain valuable insights into patients' experiences, emotions, and satisfaction levels, which are crucial for improving the quality of healthcare delivery. However, manually analysing such vast amounts of audio data is time-consuming and impractical. This is where automated sentiment analysis techniques come into play, offering a promising solution to extract meaningful insights from audio recordings efficiently.

1.2 Background & Motivation

Audio sentiment analysis, a subfield of natural language processing (NLP), focuses on automatically detecting and interpreting emotions, attitudes, and sentiments expressed in spoken language. While sentiment analysis has been extensively studied in text data, its application to audio recordings presents unique challenges due to variability in speech patterns, accents, background noise, and other acoustic factors. Despite these challenges, recent advancements in machine learning algorithms and audio processing techniques have made audio sentiment analysis increasingly feasible and accurate.

1.3 Objective

The objective of the research is to conduct sentiment analysis on unlabeled audio recordings of medical conversations between patients and healthcare providers. Initially, you will perform shallow sentiment analysis on transcriptions of the audio data to understand the sentiment expressed, identifying instances of both negative and positive sentiment. Subsequently, we will be applying deep learning models, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), to further analyze the sentiment of the audio recordings. This will involve extracting audio features and employing LSTM and BiLSTM models for sentiment analysis. Finally, we will compare the results obtained from these deep learning models with the results from the shallow sentiment analysis to evaluate the effectiveness and performance of the deep learning approach. Additionally, applying LSTM, BiLSTM, and BERT models on text data to perform sentiment analysis, and then comparing the results obtained by these models. This comprehensive approach aims to leverage both audio and text data to gain insights into the sentiment expressed in medical conversations.

1.4 Contribution

In this study, unlabelled data comprising both audio recordings and transcriptions of medical conversations were subjected to shallow sentiment analysis, initially focusing on the textual transcriptions. Leveraging established rules of text preprocessing, including tasks such as removing prefixes, punctuation, spelling correction, stemming, and lemmatization, the transcriptions were refined to enhance the accuracy of subsequent sentiment analysis. Following this preprocessing stage, sentiment analysis was performed on the transcriptions, providing preliminary insights into the sentiment dynamics of the conversations. Subsequently, these sentiment scores were mapped back to their respective audio files, laying the groundwork for deeper sentiment analysis.

Building upon the insights gained from the shallow sentiment analysis, deep learning models were employed to analyze the sentiment of both the audio recordings and the textual transcriptions. Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models were applied to the audio files, leveraging their capabilities in capturing temporal dependencies within sequential data. Concurrently, BERT (Bidirectional Encoder Representations from Transformers) models were utilized on the textual transcriptions to extract contextualized representations and discern sentiment nuances. By integrating both shallow and deep learning techniques across audio and textual data modalities, this research aims to provide a comprehensive understanding of sentiment expression in medical conversations, contributing to improved patient care and communication in clinical settings.

1.5 Outline of the report

In the report, we are conducting sentiment analysis on unlabelled audio recordings of medical conversations between patients and healthcare providers. The methodology involves several stages: first, employing shallow sentiment analysis techniques on transcriptions of the audio data to gain initial insights into the expressed sentiment. Subsequently, mapping these sentiments to their respective audio files facilitates the application of sentiment analysis using Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models. Additionally, sentiment analysis is performed directly on the transcripts using LSTM, BiLSTM, and BERT models, which are advanced deep learning techniques. Finally, the study compares the results obtained from these various approaches to evaluate their effectiveness and performance in analyzing sentiment in medical conversations.

CHAPTER 2: LITERATURE SURVEY

Sentiment analysis, a key task in natural language processing (NLP), has traditionally focused on textual data. However, with the advent of multimodal data, integrating textual and audio inputs has gained prominence for achieving more nuanced sentiment analysis. This survey explores the evolution of models and techniques that have laid the groundwork for CM-BERT, a cross-modal BERT model designed for text-audio sentiment analysis.

Textual Sentiment Analysis

Traditional Approaches: Early methods for sentiment analysis primarily relied on machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. These models used handcrafted features and bag-of-words representations, which often failed to capture the contextual nuances of language [1].

Deep Learning Models: The introduction of deep learning revolutionized sentiment analysis. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, became popular due to their ability to manage sequential data and capture temporal dependencies [2]. Bidirectional LSTMs (BiLSTMs) further enhanced performance by processing input data in both forward and backward directions, thus understanding context more effectively [3].

Transformer Models: The advent of transformer-based models, especially BERT (Bidirectional Encoder Representations from Transformers), marked a significant leap in NLP. BERT's deep bidirectional nature allows it to understand the context of a word based on all its surroundings, making it exceptionally powerful for sentiment analysis [4]. Studies have consistently shown BERT outperforming LSTM and BiLSTM models, achieving state-of-the-art results in various NLP tasks, including sentiment analysis [4].

Audio Sentiment Analysis

Feature Extraction: Audio sentiment analysis involves extracting features such as pitch, tone, and energy from audio signals. Traditional methods utilized handcrafted features and statistical models [5].

Deep Learning Approaches: With the rise of deep learning, models such as Convolutional Neural Networks (CNNs) and LSTMs were adapted for audio sentiment analysis. These models improved the ability to capture the emotional content in audio signals. Specifically, BiLSTM models demonstrated superior performance by effectively capturing the temporal dynamics of audio features, surpassing standard LSTM models with higher accuracy, F1 scores, precision, and recall [6].

CHAPTER 3: METHODOLOGY

3.1 Problem Statement

This research aims to overcome the challenge of sentiment analysis in unlabelled audio recordings of medical conversations, where the absence of labelled sentiment data impedes traditional analysis methods. Leveraging both shallow and deep learning approaches, the study seeks to develop a methodology for sentiment analysis in such recordings.

While shallow sentiment analysis on transcriptions provides initial insights, it may lack depth in capturing nuanced sentiment expressions. Hence, the research explores the effectiveness of deep learning models like Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) in directly extracting sentiment from audio signals. By addressing this challenge, the study endeavours to advance sentiment analysis techniques for healthcare applications, contributing to enhanced patient care and communication in clinical settings.

3.2 Proposed Techniques / Approach

In this proposed technique, the sentiment analysis process begins with the application of shallow sentiment analysis to the transcriptions of medical conversations. This initial step involves employing rules of text preprocessing to refine the transcriptions for enhanced sentiment analysis accuracy this was achieved by using the Spacy library available in python. Following this, sentiment which are either positive or negative derived from the transcriptions are mapped back to their corresponding audio files, establishing a link between textual sentiment expressions and audio signals. Subsequently, deep learning models, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), are applied to the audio files. MFCC (Mel-Frequency Cepstral Coefficients) features are extracted from the audio data to serve as input for the LSTM and BiLSTM models, enabling the capture of sentiment expression within the audio recordings.

In parallel, sentiment analysis is conducted on the textual transcriptions using a combination of deep learning models. BERT (Bidirectional Encoder Representations from Transformers) models are employed to extract contextualized representations from the transcriptions, capturing intricate nuances of sentiment expression. Additionally, LSTM and BiLSTM models are applied to the transcriptions, leveraging their capabilities in processing sequential data to further refine the sentiment analysis results. By integrating both shallow and deep learning techniques across audio and textual data modalities, this proposed technique aims to provide a comprehensive understanding of sentiment dynamics in medical conversations, ultimately contributing to improved patient care and communication in clinical settings.

3.2.1 Audio Processing

Audio feature extraction is a critical preprocessing step in audio signal processing tasks, including speech recognition, music classification, and sound event detection. It involves transforming raw audio waveforms into a set of numerical features that capture relevant information about the audio signal. These features are then used as input to machine learning models for further analysis and classification.

One commonly used technique for audio feature extraction is Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are a representation of the short-term power spectrum of a sound, modelled after the human auditory system's response to sound. The process involves several steps:

1. **Frame Segmentation:** The audio signal is divided into short overlapping frames, typically 20-30 seconds in duration but for this experiment, the total duration was taken 420 seconds considering on an average looking at the audio files used for this experiment. This segmentation allows for analysis of the audio signal over short time intervals, capturing temporal dynamics.
2. **Pre-emphasis:** Pre-emphasis is applied to the signal to amplify high-frequency components, enhancing the signal-to-noise ratio and improving feature extraction accuracy.
3. **Windowing:** Each frame is multiplied by a window function, such as the Hamming window, to reduce spectral leakage and improve frequency resolution.
4. **Fast Fourier Transform (FFT):** The Fourier transform is applied to each windowed frame to obtain its frequency-domain representation. This process converts the audio signal from the time domain to the frequency domain, revealing the frequency content of the signal.
5. **Mel Filterbank:** A set of triangular filters, spaced uniformly in the Mel frequency scale, is applied to the power spectrum obtained from the FFT. This step mimics the non-linear frequency resolution of the human auditory system.
6. **Logarithmic Compression:** The logarithm of the filterbank energies is computed to compress the dynamic range of the spectrum. This step improves the discriminative power of the features and makes them more robust to variations in signal intensity.
7. **Discrete Cosine Transform (DCT):** Finally, the DCT is applied to the log-filterbank energies to decorrelate the features and reduce dimensionality. The resulting coefficients, known as MFCCs, capture essential spectral characteristics of the audio signal and serve as representative features for further analysis.

Overall, audio feature extraction plays a crucial role in analysing and understanding audio signals, enabling us to do the experiment on the audio data for sentiment analysis

Audio data that was used for sentiment analysis was unlabelled, having the audio transcriptions in hand, we did a shallow sentiment analysis on the transcription to obtain the sentiment for each audio files.

3.2.2 Text Processing

The transcription of the audio files underwent preprocessing using rules defined in the NLTK and spaCy libraries. This included steps such as removing prefixes, punctuation, and performing stemming and lemmatization to normalize the text.

Following the preprocessing of the transcripts, shallow sentiment analysis was applied using the VADER (Valence Aware Dictionary and sentiment Reasoner) sentiment analysis tool from NLTK. VADER is specifically designed to analyse sentiment in text data and provides polarity scores indicating the positivity, negativity, and neutrality of the text.

we focused on classifying sentiment as either positive or negative, disregarding neutral sentiment due to its minimal occurrence in the dataset. After obtaining sentiment labels for the transcripts, they were mapped to their corresponding audio files.

3.2.3 Models: LSTM, BiLSTM, BERT

LSTM

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. Its architecture includes memory cells and gates that control the flow of information, allowing it to selectively update and forget information over time.

Advantages:

1. Long-Term Dependencies: Capable of capturing long-range dependencies in data.
2. Gating Mechanisms: Uses gates to regulate information flow, addressing the vanishing gradient problem.
3. Flexibility: Highly adaptable to various tasks by adjusting architecture and parameters.

Disadvantages:

1. Complexity: More complex architecture compared to traditional RNNs.
2. Computationally Intensive: Requires more computational resources, leading to longer training times.
3. Overfitting: Susceptible to overfitting, especially with small datasets or complex architectures.
4. Training Difficulty: Training can be challenging, especially with long sequences or noisy data, requiring careful tuning of hyperparameters and training strategies.

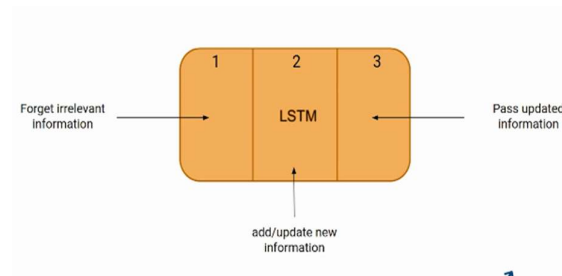


Fig 1: LSTM network architecture¹

1. ¹ <https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/>

BiLSTM

BiLSTM (Bidirectional Long Short-Term Memory) is an extension of LSTM that processes input data in both forward and backward directions. It captures contextual information from past and future inputs, enabling it to better understand sequential data and make more accurate predictions.

Advantages:

1. Contextual Understanding: Captures information from both past and future inputs, enhancing understanding of sequential data.
2. Improved Performance: Often yields better performance than traditional LSTM by considering a broader context.
3. Versatility: Adaptable to various tasks in natural language processing, speech recognition, and more.

Disadvantages:

1. Increased Complexity: More complex architecture compared to LSTM, potentially leading to longer training times and higher computational costs.
2. Overfitting: Susceptible to overfitting, especially with large and complex datasets.
3. Training Difficulty: Requires careful hyperparameter tuning and regularization to prevent overfitting and ensure optimal performance.

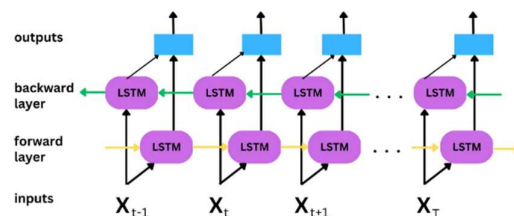


Fig 2: BiLSTM Architecture²

2. ² <https://medium.com/@souro400.nath/why-is-bilstm-better-than-lstm-a7eb0090c1e4>

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing (NLP) model developed by Google. It utilizes a transformer architecture and pretraining on large text corpora to generate contextualized word embeddings, allowing it to understand the meaning of words in context.

Advantages:

1. **Contextual Understanding:** Captures contextual information from both left and right contexts, improving understanding of language nuances.
2. **Pretraining:** Pretrained on large text corpora, enabling transfer learning and adaptation to various downstream NLP tasks.
3. **State-of-the-Art Performance:** Achieves impressive performance on a wide range of NLP tasks, surpassing previous models in accuracy and efficiency.

Disadvantages:

1. **Computational Resources:** Requires substantial computational resources for pretraining and fine-tuning, making it less accessible for small-scale projects.
2. **Complexity:** More complex architecture compared to traditional NLP models, potentially leading to longer training times and increased model size.
3. **Domain Specificity:** Performance may vary across different domains and languages, requiring fine-tuning or adaptation for optimal results.

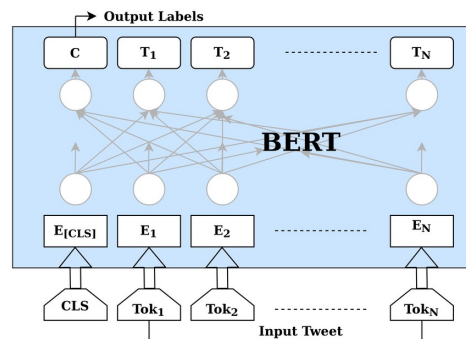


Fig 3: BERT architecture³

3. ³ https://www.researchgate.net/figure/BERT-model-architecture_fig2_348214408

CHAPTER 4: EXPERIMENTS & RESULTS

4.1 Dataset

The dataset is divided into two sets of files: audio files of the simulated conversations in mp3 format, and the transcripts of the audio files as text files. There are 272 mp3 audio files and 272 corresponding transcript text files. Each file is titled with three characters and four digits. RES stands for respiratory, GAS represents gastrointestinal, CAR is cardiovascular, MSK is musculoskeletal, DER is dermatological, and the four following digits represent the case number of the respective disease category. This data set was used for sentiment analysis.

4.1.1 Dataset Statistics

Distribution by Disease Category

Distribution of Disease Category is shown in Table 1.

Disease Category	No. of Transcripts	No. of Audio
RES	213	213
GAS	6	6
CAR	5	5
MSK	46	46
DER	1	1

Table 1: Distribution by Disease Category

Disease Category by Sentiment

Distribution of Disease category categorized into two sentiments either positive or negative is shown in Table 2.

Disease Category	Positive	Negative
RES	202	11
GAS	5	1
CAR	5	0
MSK	37	9
DER	1	0

Table 2: Disease Category by Sentiment

Processed Dataset

Dataset used after initial shallow sentiment is shown in Table 3.

Audio File	Transcript File	Transcripts	Sentiment
CAR0001_positive.mp3	CAR0001.txt	D: What brought you in today? P: Sure, I'm I'm...so I thought I should get it checked out.. . . P: No,...me everything.	Positive
RES0207_positive.mp3	RES0207.txt	D: Well can...you in today? . . P: Yeah, thats great. D: OK. Sure see you shortly.	Positive
GAS0001_positive.mp3	GAS0001.txt	D: What brings you in today? P: I've just...right now. . . D: You're welcome.	Positive
MSK0024_negative.mp3	MSK0024.txt	D: What brings you here today? . . D: You're welcome, alright take care.	negative

Table 3: Snapshot of the dataset

4.2 Experimental setup

4.2.1 Audio

Preprocessing: For preprocessing text data before sentiment analysis. It utilizes various libraries such as NLTK, spaCy, and regular expressions to perform tasks like tokenization, stopword removal, punctuation removal, stemming, and lemmatization. The `preprocess_text` function combines these preprocessing steps into a single pipeline, which can be applied to each text input. Additionally, there's a `sentiment_analysis` function using VADER (Valence Aware Dictionary and sentiment Reasoner) for sentiment analysis. The `preprocess_files` function reads text files, preprocesses the text using `preprocess_text`, performs sentiment analysis, and writes the results along with file metadata to a CSV file. Finally, an example usage section demonstrates how to preprocess a list of text files and output the results to a CSV file while renaming corresponding audio files based on sentiment analysis results.

Example:

Here we are taking the raw input sting then applying the rules of text pre-processing to get the pre-processed string on which we are applying the shallow sentiment.

Transcript	Raw text	Pre-processed text
RES0010.txt	D: How may I help you? . . . P: I just came in here today, I've been having a lot of...recently. Just in my left shoulder.	how may help you...today lot shoulder pain kind left
CAR0002.txt	D: What brings you in here today . . . P: It started just 30 minutes ago.	yea thi pain chest ok is the pain exactly...left thi pain

Tabel 3: Snapshot of the Pre-processing

In Table 2, after the text are pre-processed, we apply shallow sentiment analysis on it which is shown in Table 1 which is used for Sentiment analysis using Deep Learning Models.

General Setup

- **Data Preprocessing:** Each model begins by loading preprocessed data from CSV files containing text data and corresponding sentiment labels. Text preprocessing involves tokenization using either BERT tokenizer or Keras Tokenizer, and padding sequences to ensure uniform length.

- **Training and Evaluation:** All models utilize a common procedure for dataset splitting, with 80% of the data allocated for training and 20% for testing. Evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to assess model performance.

Model-Specific Configurations

LSTM

- **Architecture:** The LSTM model architecture consists of an Embedding layer followed by a SpatialDropout1D layer for regularization. A single LSTM layer with 196 units is incorporated to capture long-term dependencies in the text data, followed by a Dense layer with softmax activation for sentiment classification.
- **Hyperparameters:** Hyperparameters such as the maximum number of features, dropout rates, and LSTM units are specified for the LSTM model.
- **Training Details:** The LSTM model is trained for 10 epochs with a batch size of 128, and validation data is used to monitor model performance during training.

BiLSTM

- **Architecture:** The BiLSTM model architecture differs from LSTM by incorporating a Bidirectional LSTM layer, enabling it to learn from both forward and backward sequences. Apart from this, the architecture remains similar to LSTM, with embedding, dropout, and dense layers.
- **Hyperparameters:** Hyperparameters for the BiLSTM model are similar to those of LSTM, with adjustments made for bidirectional processing.
- **Training Details:** Training of the BiLSTM model follows a similar procedure to LSTM, including dataset splitting, epoch specification, and validation monitoring.

4.2.2 Text

Preprocessing: The audio files were initially in the MP3 format but were converted to the WAV format. This conversion was necessary as some audio processing libraries and models may require WAV format due to its uncompressed nature, which ensures better quality and compatibility with various audio processing tasks

General Setup

- **Data Preprocessing:** Each model begins by loading preprocessed data from CSV files containing text data and corresponding sentiment labels. For BERT and BiLSTM models, specific columns ('Text' and 'Sentiment') are selected from the DataFrame, while for LSTM, the entire DataFrame is filtered to include only these two columns. Text tokenization and padding are applied to convert text sequences into suitable input formats for each model.
- **Training and Evaluation:** All models utilize a train-test split of 80% for training and 20% for testing. DataLoader objects are created to efficiently load data in batches during model training and evaluation. Model evaluation includes computation of classification metrics such as precision, recall, F1-score, and accuracy using scikit-learn functions.

Model-Specific Configurations

BERT

- **Loading and Preprocessing Data:** CSV file containing text data is loaded and relevant columns are selected. BERT tokenizer is employed for tokenization, and sentiment labels are mapped to numerical values. Data is split into training and testing sets, and DataLoader objects are created.
- **Model Initialization:** Pre-trained BERT model (BertForSequenceClassification) is loaded, followed by initialization of optimizer and scheduler.
- **Training and Evaluation:** Custom train function is defined for training the BERT model, and an evaluate function is defined for model evaluation on test data.

LSTM

- **Data Loading and Preprocessing:** Similar to BERT, CSV file is loaded, and relevant columns are selected. Text tokenization and padding are applied using Keras Tokenizer and pad_sequences.
- **Model Initialization:** Sequential model is initialized with an Embedding layer, SpatialDropout1D layer, LSTM layer, and Dense layer with softmax activation.
- **Training and Evaluation:** Model is compiled with categorical cross-entropy loss and Adam optimizer. Training is performed for 10 epochs with batch size of 128, followed by evaluation on test data.

BiLSTM

- **Data Preparation:** Similar to LSTM, CSV file is loaded, relevant columns are selected, and text tokenization and padding are applied.
- **Model Building:** Sequential model is initialized with additional Bidirectional LSTM layer for learning long-term dependencies in both forward and backward directions.
- **Training and Evaluation:** Model is trained and evaluated using similar procedures as LSTM, with data splitting, compilation, training, and evaluation steps.

4.3 Results

4.3.1 Sentiment analysis

Here we have performed experiment based on Audio files as well on Transcripts. We have used 80% of the dataset for training purpose and the remaining 20% for testing purpose. We notice that BiLSTM performs best, with an accuracy of 87%, precision score 0.87, recall score 0.87 and f1 score 0.87 on Audio dataset and BERT performs best with 91% accuracy with precision score of 0.91, recall of 0.91, F1 score of 0.91.

The detail results are shown below:

		Audio		Text		
		LSTM	BiLSTM	LSTM	BiLSTM	BERT
Overall	Precision	0.69	0.87	0.74	0.75	0.91
	Recall	0.69	0.87	0.69	0.67	0.91
	F1 score	0.68	0.87	0.66	0.61	0.91
	Accuracy	69%	87%	65%	71%	91%
Positive	Precision	0.69	0.84	0.67	0.65	0.88
	Recall	0.80	0.88	0.91	0.95	0.96
	F1 score	0.74	0.86	0.77	0.77	0.92
	Accuracy	69%	84%	67%	65%	88%
Negative	Precision	0.69	0.90	0.50	0.85	0.94
	Recall	0.55	0.86	0.17	0.37	0.84
	F1 score	0.61	0.88	0.26	0.52	0.89
	Accuracy	69%	90%	50%	85%	94%

Table 4: Sentiment Analysis result

4.4 Error Analysis

4.4.1 LSTM

Audio

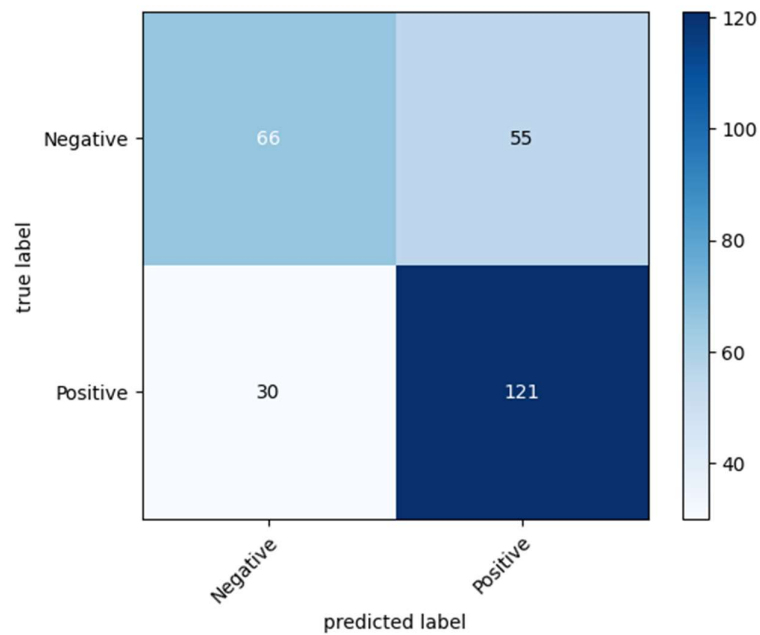


Fig 4: Confusion Matrix LSTM (Audio)

Here is the interpretation:

- **Negative (Class 0):** 66 audio dataset were correctly classified (True Negative) & 55 audio dataset were incorrectly classified (False Positive) by our model.
- **Positive (Class 1):** 121 audio dataset were correctly classified (True Positive) & 30 audio dataset were incorrectly classified (False Negative) by our model.

Text

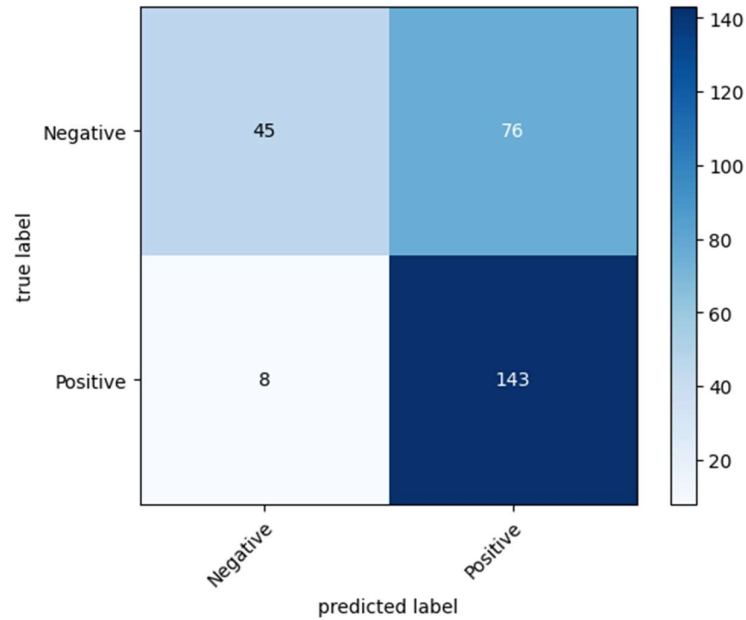


Fig 5: Confusion Matrix LSTM (Text)

Here is the interpretation:

- **Negative (Class 0):** 45 text dataset were correctly classified (True Negative) & 76 text dataset were incorrectly classified (False Positive) by our model.
- **Positive (Class 1):** 8 text dataset were correctly classified (False Negative) & 143 audio dataset were incorrectly classified (True Positive) by our model.

Table 4 shows some of the examples that were misclassified by LSTM

mode	Actual tag	Predicted tag
RES0075 (Audio)	positive	negative
GAS0002 (Audio)	negative	positive
CAR0001 (Text transcript from raw Data set)	positive	negative
MSK0045 (Text transcript from raw Data set)	negative	positive

Table 5: LSTM prediction error

4.4.2 BiLSTM

Audio

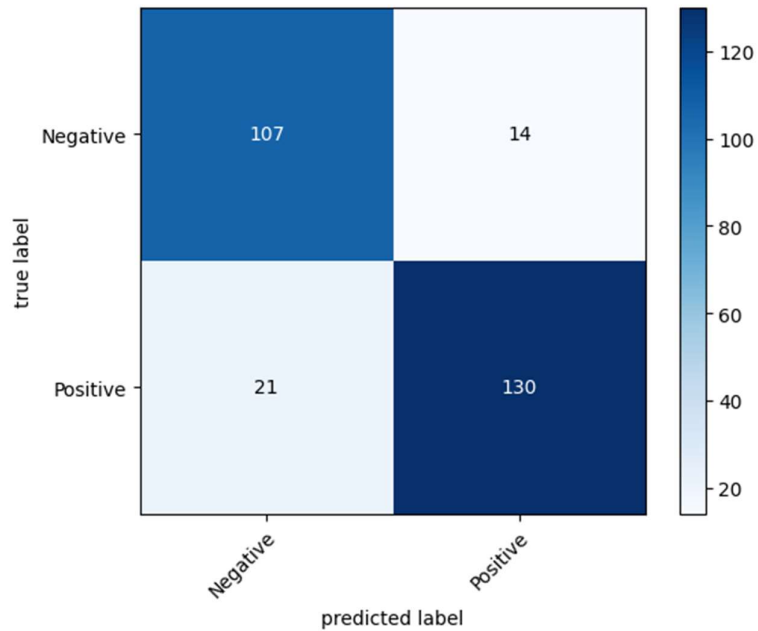


Fig 6: confusion matrix BiLSTM (Audio)

Here is the interpretation:

- **Negative (Class 0):** 107 text dataset were correctly classified (True Negative) & 14 text dataset were incorrectly classified (False Positive) by our model.
- **Positive (Class 1):** 21 text dataset were correctly classified (False Negative) & 130 audio dataset were incorrectly classified (True Positive) by our model

Text

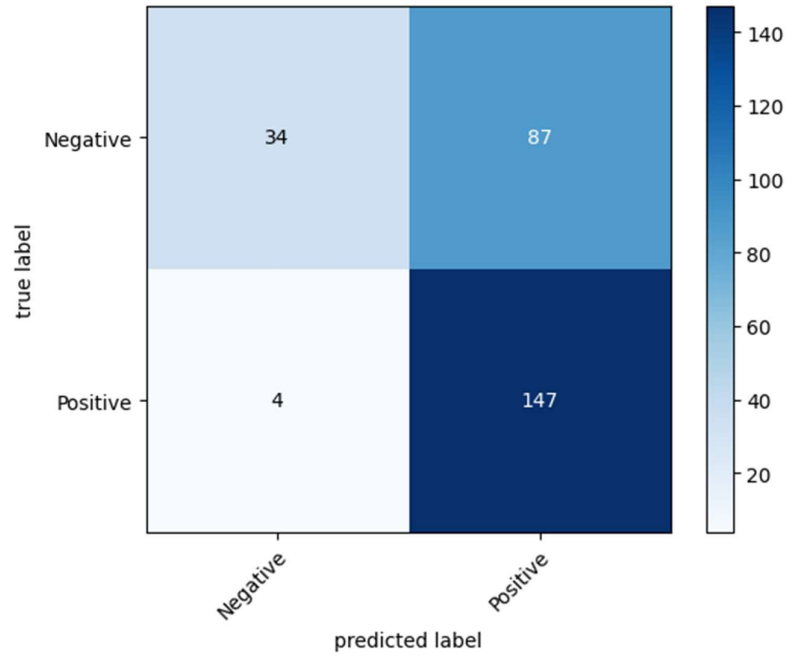


Fig 7: Confusion matrix BiLSTM (Text)

Here is the interpretation:

- **Negative (Class 0):** 34 text dataset were correctly classified (True Negative) & 87 text dataset were incorrectly classified (False Positive) by our model.
- **Positive (Class 1):** 4 text dataset were correctly classified (False Negative) & 147 audio dataset were incorrectly classified (True Positive) by our model

Table 5 shows some of the examples that were misclassified by BiLSTM

mode	Actual tag	Predicted tag
RES0208(Audio)	positive	negative
MSK0009(Audio)	negative	positive
RES0073(Text transcript from raw Data set)	positive	negative
GAS0005(Text transcript from raw Data set)	negative	positive

Table 6: BiLSTM prediction error

4.4.3 BERT

Text

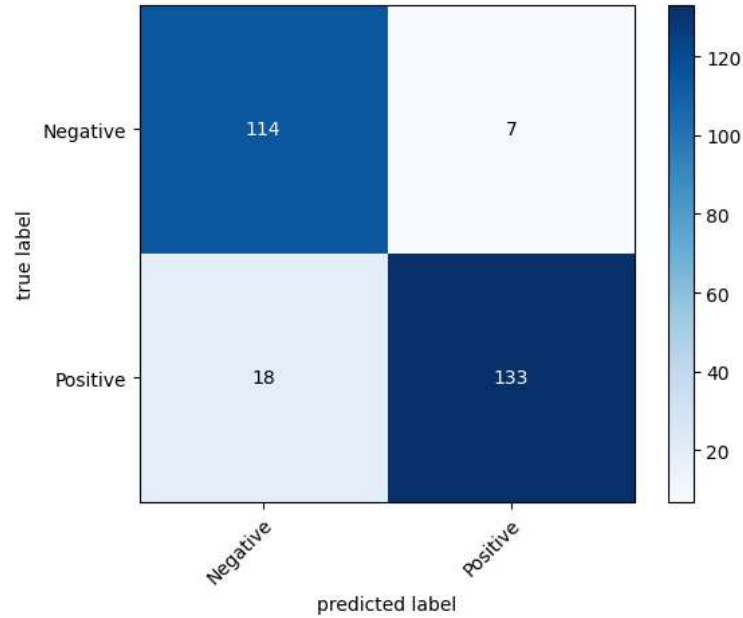


Fig 8: Confusion Matrix BERT (Text)

Here is the interpretation:

- **Negative (Class 0):** 114 text dataset were correctly classified (True Negative) & 7 text dataset were incorrectly classified (False Positive) by our model.
- **Positive (Class 1):** 18 text dataset were correctly classified (False Negative) & 133 audio dataset were incorrectly classified (True Positive) by our model

Table 6 shows some of the examples that were misclassified by BERT

Mode	Actual tag	Predicted tag
RES0072(transcript)	positive	negative
MSK0029(transcript)	negative	positive

Table 7: BERT prediction error

4.5 Discussion

In text sentiment analysis, BERT has shown a significant improvement over traditional LSTM and BiLSTM models. BERT achieved an impressive overall accuracy of 91%, with an F1 score, precision, and recall all at 0.91. The advantage of BERT over LSTM and BiLSTM lies in its ability to understand context and semantics more effectively through its bidirectional training and attention mechanisms, which allows it to capture nuanced meanings in text.

BERT in Text Sentiment Analysis

Good Example:

- **Text:** "I have this pain in my chest. It's just right over on the left side. It started just 30 minutes ago while I was shovelling the driveway. The pain has gotten worse ever since it started. It feels dull, like there's a lot of pressure on my chest."
- **Sentiment Prediction:** Negative
- **Performance:** BERT accurately identifies the negative sentiment due to its ability to understand the context and severity of the situation. Words like "pain," "worse," and "pressure" are strong indicators of a negative sentiment, which BERT correctly captures.

Bad Example:

- **Text:** "I live in a house with my parents. I drive a bus for the city. Usually on Sundays I'll go for a run."
- **Sentiment Prediction:** Neutral
- **Performance:** BERT might misclassify the sentiment if the text does not contain clear indicators of positive or negative sentiment. In this example, the text is largely neutral and lacks strong emotional cues, making it challenging for BERT to provide a precise sentiment classification.

In contrast, for audio sentiment analysis, BiLSTM demonstrated notable superiority over the traditional LSTM model, attaining an overall accuracy of 87%, with matching F1 score, precision, and recall of 0.87. The advantage of BiLSTM over LSTM in audio data processing is its ability to capture dependencies from both past and future contexts in the audio sequence, leading to a more comprehensive understanding of the sentiment conveyed in the audio recordings.

BiLSTM in Audio Sentiment Analysis

Good Example:

- **Audio Description:** A patient describes their chest pain with a tense and distressed tone, speaking rapidly and with a higher pitch, indicating discomfort and urgency.
- **Sentiment Prediction:** Negative
- **Performance:** BiLSTM effectively captures the emotional cues and temporal dependencies in the audio, leading to an accurate prediction of negative sentiment. The model can detect the stress and discomfort in the patient's voice, which are clear indicators of negative sentiment.

Bad Example:

- **Audio Description:** A patient describes their daily routine in a calm and steady tone, with no significant changes in pitch or pace.
- **Sentiment Prediction:** Neutral

- **Performance:** BiLSTM might misclassify the sentiment if the emotional cues in the audio are subtle or if the speaker maintains a neutral tone throughout. In this case, the lack of clear emotional inflection makes it harder for the model to accurately detect any sentiment, leading to a potential misclassification.

These results highlight the advanced capabilities of BERT for text analysis and the effectiveness of BiLSTM for audio data.

CHAPTER 5: CONCLUSION & FUTURE WORK

Conclusion:

In this study, we investigated sentiment analysis techniques applied to medical conversations, focusing on both audio recordings and their corresponding transcriptions. Shallow sentiment analysis provided initial insights into the sentiment expressed in the transcribed text, which was then mapped to the respective audio files. Subsequently, deep learning models, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), were applied to analyze the sentiment of the audio recordings. Furthermore, sentiment analysis was performed on the transcriptions using LSTM, BiLSTM, and BERT models. By comparing the results of these techniques, we gained valuable insights into their effectiveness and performance in the context of medical dialogue. The findings suggest the potential of deep learning models, particularly BiLSTM, in accurately capturing sentiment from audio data. Additionally, the comparison with shallow sentiment analysis highlighted the advantages of deep learning approaches in handling complex medical conversations.

Future Work:

There are several avenues for future research based on the findings of this study. Firstly, exploring alternative deep learning architectures, such as transformer-based models like BERT, could provide further improvements in sentiment analysis accuracy for both audio and text data. Additionally, investigating multimodal approaches that integrate both audio and text features may enhance the robustness and generalizability of sentiment analysis in medical conversations. Expanding the dataset to include a broader range of medical contexts or languages would also contribute to the understanding of sentiment expression in diverse healthcare settings. Furthermore, addressing any identified limitations and refining the methodologies used in this study could lead to more comprehensive sentiment analysis frameworks tailored specifically for medical dialogue data.

REFERENCES

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
3. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.
5. Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. *ACM Multimedia*.
6. Lee, S., & Han, C. (2020). Bidirectional LSTM for Audio Sentiment Analysis. *IEEE Access*.