

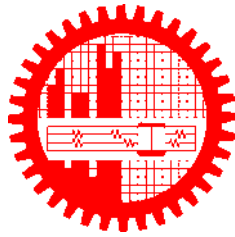
M.Sc. Engg. (CSE) Thesis

A Deep Ensemble Approach of Anger Detection From Audio-Textual Conversation

Submitted by
Mahjabin Nahar
1017052003

Supervised by
Dr. Md. Shamsuzzoha Bayzid

Co-Supervised by
Dr. Mohammed Eunos Ali



Submitted to
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science and Engineering

May 2022

Candidate's Declaration

I, do, hereby, certify that the work presented in this thesis, titled, "A Deep Ensemble Approach of Anger Detection From Audio-Textual Conversation", is the outcome of the investigation and research carried out by me under the supervision of Dr. Md. Shamsuzzoha Bayzid, Associate Professor, and the co-supervision of Dr. Mohammed Eunus Ali, Professor, Department of CSE, BUET.

I also declare that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

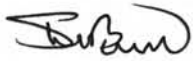



Mahjabin Nahar

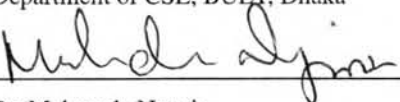
1017052003


The thesis titled “A Deep Ensemble Approach of Anger Detection From Audio-Textual Conversation”, submitted by Mahjabin Nahar, Student ID 1017052003, Session October 2017, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfilment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents on May 15, 2022.

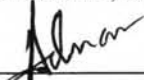
Board of Examiners

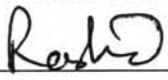
1. 

Dr. Md. Shamsuzzoha Bayzid
Associate Professor
Department of CSE, BUET, Dhaka
Chairman
(Supervisor)
2. 

Dr. Mohammed Eunos Ali
Professor
Department of CSE, BUET, Dhaka
Member
(Co-Supervisor)
3. 

Dr. Mahmuda Naznin
Professor and Head
Department of CSE, BUET, Dhaka
Member
(Ex-Officio)
4. 

Dr. Rifat Shahriyar
Professor
Department of CSE, BUET, Dhaka
Member
5. 

Dr. Muhammad Abdullah Adnan
Associate Professor
Department of CSE, BUET, Dhaka
Member
6. 

Dr. Mohammad Rashedur Rahman
Professor
Department of Electrical and Computer Engineering
North South University, Dhaka
Member
(External)

Acknowledgement

I am thankful to the Almighty for the health and well-being that He bestowed upon me.

I wish to thank my thesis supervisor Dr. Mohammed Eunos Ali for his continuous support and guidance throughout the years, without which this work would never have been completed. I would also like to thank my present thesis supervisor Dr. Md. Shamsuzzoha Bayzid for his support and assistance. Moreover, I would like to express my gratitude to Dr. Mahmuda Naznin, Head of the Department, CSE, BUET, the department faculty members, and colleagues for their continuous encouragement and support.

I must thank my husband, Nafis Irtiza Tripto, for his continuous support and encouragement during my thesis. I am also indebted to my parents for their unceasing support and encouragement. Finally, I would like to thank my son Numayir Mehraan for being the sweetest baby ever and letting me work on my thesis.

Dhaka
May 15, 2022

Mahjabin Nahar
1017052003

Contents

Candidate’s Declaration	i
Board of Examiners	ii
Acknowledgement	iii
List of Figures	vii
List of Tables	viii
List of Algorithms	x
Abstract	xi
1 Introduction	1
1.1 Motivation	1
1.2 Background	2
1.3 State of the Art	3
1.3.1 Research in Offline Anger Detection	3
1.3.2 Research in Online Anger Detection	3
1.3.3 Research in the Bengali Language	4
1.4 Problem and Overview of Solution	4
1.5 Research Objective	6
1.6 Contribution	6
1.7 Organization	7
2 Background	8
2.1 Features used in Audio-Textual Speech Processing	8
2.1.1 Audio Features	8
2.1.2 Textual Features	9
2.2 Models and Techniques used in Audio-Textual Speech Processing	10
2.2.1 Traditional Machine Learning Models	10
2.2.2 Deep Learning Models	11

2.2.3	Ensemble Machine Learning Models	14
3	Related Work	16
3.1	Research in Speech Communication	16
3.2	Research in Emotion Recognition	16
3.2.1	Emotion Recognition from Speech	16
3.2.2	Emotion Recognition from Text	17
3.2.3	Multimodal Emotion Recognition	17
3.3	Research in Offline Anger Recognition	17
3.3.1	Offline Anger Recognition across Different Languages	18
3.3.2	Offline Anger Recognition across Different Modalities and Feature Sets	18
3.3.3	Offline Anger Recognition Incorporating Different Techniques	19
3.4	Research in Online Anger Recognition	19
3.5	Research in the Bengali Language	20
3.5.1	Emotion Recognition from Bengali Speech	20
3.5.2	Emotion Recognition from Bengali Text	21
4	Offline Anger Detection	22
4.1	Methodologies	22
4.1.1	Intuitions behind the Proposed Approach	22
4.1.2	Feature Extraction	23
4.1.3	Feature Extraction from Audio	23
4.1.4	Feature Extraction from Text	24
4.1.5	Offline Anger Detection	25
4.2	Experimental Evaluation	29
4.2.1	Dataset Description	29
4.2.2	Baseline Methods: Feature Sets and Fusion Combinations	31
4.2.3	Baseline Methods: Classifiers	33
4.2.4	Baseline Methods: Fusion Classifiers	35
4.3	Results and Discussions	35
4.3.1	Results: Bengali Call-Center Dataset	35
4.3.2	Results: IEMOCAP Dataset	37
4.3.3	Discussions and Limitations	38
5	Online Anger Detection	40
5.1	Methodologies	40
5.1.1	Intuitions behind the Proposed Approach	40
5.1.2	Modeling the Input and Output Data	41
5.1.3	Feature Extraction	43

5.1.4	Online Anger Detection	43
5.2	Experimental Evaluation	45
5.2.1	Dataset Description	45
5.2.2	Baseline Methods: Feature Sets and Fusion Combinations	46
5.2.3	Baseline Methods: Classifiers	47
5.2.4	Baseline Methods: Fusion Classifiers	48
5.3	Results and Discussions	48
5.3.1	Results: Bengali Call-Center Dataset	48
5.3.2	Results: IEMOCAP Dataset	52
5.3.3	Discussions and Limitations	56
6	Conclusion	57
	References	58

List of Figures

1.1	An overview of Offline and Online Anger Detectors	4
1.2	A detailed overview of the Offline Anger Detector, OAD and the Online Anger Detector, EAD	5
2.1	The architecture of a Bidirectional-LSTM (BiLSTM) network	12
2.2	The architecture of a Convolutional Neural Network (CNN)	13
2.3	The architecture of a Transformer network	14
4.1	Traditional or Offline Anger Detector with Gender Classifier	26
4.2	Gender Classifier	26
4.3	Traditional or Offline Anger Detector, OAD	28
5.1	Input utterances, output utterances, and skipped utterances explained	42
5.2	Early or Online Anger Detector, EAD	45
5.3	Impact of varying the number of skipped utterances, s , on the F1 score of online anger for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset	50
5.4	Impact of varying the number of skipped utterances, s , on the F1 score of online anger for audio and textual features considering both speakers for the IEMOCAP dataset	54

List of Tables

4.1	Percentage of angry utterances in the Bengali Call-center dataset	30
4.2	Percentage of male and female speakers in the Bengali Call-center dataset	30
4.3	Percentage of angry utterances in the IEMOCAP dataset	31
4.4	Percentage of male and female speakers in the IEMOCAP dataset	31
4.5	F1 score of offline anger for different feature sets and models for the Bengali call-center dataset	36
4.6	F1 score of offline anger for different classifiers after fusion for the Bengali call-center dataset	36
4.7	F1 score of offline anger for different feature sets and models for the IEMOCAP dataset	37
4.8	F1 score of offline anger for different state-of-the-art models for the IEMOCAP dataset	38
4.9	F1 score of offline anger for different classifiers after fusion for the IEMOCAP dataset	38
5.1	F1 score of online anger for different feature sets and models for the Bengali call-center dataset ($i = 3, o = 1, s = 0$)	49
5.2	F1 score of online anger for different classifiers after fusion for the Bengali call-center dataset ($i = 3, o = 1, s = 0$)	49
5.3	Impact of the number of skipped utterances, s , on the F1 score of online anger for ($i = 3, o = 1$) for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset	50
5.4	Impact of the number of input utterances, i and the number of output utterances, o , on the F1 score of online anger for EAD for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset	51
5.5	F1 score of online anger for the Bengali call-center dataset ($i = 1, o = 1, s = 0$)	51
5.6	F1 score of online anger for the Bengali call-center dataset ($i = 1, o = 2, s = 0$)	51
5.7	F1 score of online anger for the Bengali call-center dataset ($i = 2, o = 1, s = 0$)	51
5.8	F1 score of online anger for the Bengali call-center dataset ($i = 2, o = 2, s = 0$)	52
5.9	F1 score of online anger for the Bengali call-center dataset ($i = 3, o = 2, s = 0$)	52

5.10	F1 score of online anger for different feature sets and models for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)	52
5.11	F1 score of online anger for different state-of-the-art models for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)	53
5.12	F1 score of online anger for different classifiers after fusion for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)	53
5.13	Impact of the number of skipped utterances, s , on the F1 score of online anger for ($i = 3, o = 1$) for audio and textual features considering the utterances of both speakers in dyadic conversations for the IEMOCAP dataset	54
5.14	Impact of the number of input utterances, i and the number of output utterances, o , on the F1 score of online anger for EAD for audio and textual features considering the utterances of both speakers in dyadic conversations for the IEMOCAP dataset	54
5.15	F1 score of online anger for the IEMOCAP dataset ($i = 1, o = 1, s = 0$)	54
5.16	F1 score of online anger for the IEMOCAP dataset ($i = 1, o = 2, s = 0$)	55
5.17	F1 score of online anger for the IEMOCAP dataset ($i = 2, o = 1, s = 0$)	55
5.18	F1 score of online anger for the IEMOCAP dataset ($i = 2, o = 2, s = 0$)	55
5.19	F1 score of online anger for the IEMOCAP dataset ($i = 3, o = 2, s = 0$)	55

List of Algorithms

Abstract

Anger detection from conversations has many real-life applications that include improving interpersonal communications, providing customer services, and enhancing workplace performance. Despite its numerous applications in a variety of domains, anger is one of the least studied basic human emotions. The existing works on anger detection mostly deal with audio-only data, though text transcriptions can be directly obtained from spoken conversations. In this thesis, we propose novel deep learning-based approaches for offline and online anger detection from audio-textual data obtained from real-life conversations. Offline anger detection deals with detecting anger from a pre-collected audio-textual conversation, while online anger detection predicts anger in the subsequent utterances of a conversation from the previous utterances.

For offline anger detection, we introduce an ensemble approach that combines handcrafted acoustic features, SincNet-based raw waveform features, and BERT-based textual features in a mid-level fusion scheme within an attention-based CNN architecture. In addition, the model includes a gender classifier to incorporate gender information into offline anger detection. On the other hand, for online anger detection, which predicts the anger of future conversational utterances from current (and past) utterances, we propose a transformer-based technique that combines audio and textual features in a mid-level fusion scheme, utilizing an ensemble-based downstream classifier. We demonstrate the efficacy of our proposed approaches using two data sets: the Bengali call-center data set and the IEMOCAP data set. Experimental results show that our proposed approaches outperform the state-of-the-art baselines by a significant margin. For offline anger recognition, our model achieves an F1 score of 85.5% on the Bengali call-center data set and 91.4% on the IEMOCAP data set. For online anger recognition, our model yields an F1 score of 66.9% on the Bengali call-center data set and 67.7% on the IEMOCAP data set. Additionally, we vary different utterance parameters, such as the numbers of input and output utterances and observe their effect on the performance of anger detection.

Chapter 1

Introduction

Communication is the process of transferring information, emotions, ideas, or feelings from one person to another, which is one of the most essential aspects of human existence and survival. During our day-to-day lives, we heavily use speech to communicate our emotions with others. Therefore, emotion recognition has received substantial attention in speech communication research from both industry and academia [1–4]. Nevertheless, while the feeling of anger is practically universal, it is one of the least studied of the basic emotions [5]. Moreover, anger detection has a wide range of applications in a multitude of domains [6, 7]. Thus, the analysis and study of anger detection are of utmost importance. Besides, anger can be readily identified through analyzing various verbal and nonverbal expressions [8].

Therefore, in this study, we aim to perform anger recognition from audio-textual data, since text transcriptions aid in the detection of conversational anger. We focus on improving traditional or offline anger detection with the help of deep learning ensemble techniques and gender information. In addition, we introduce online or early detection of anger from audio-textual conversations.

1.1 Motivation

Anger is a negative emotional state which has a substantial influence on human actions and can often give rise to dysfunctional behavior. When left unchecked, it might lead to a plethora of interpersonal problems [9]. Anger recognition has several applications in a multitude of domains. It can be used to evaluate customer satisfaction and to monitor the behavior of call center personnel to inform higher authorities, which can enhance customer service [6, 10]. Besides, monitoring anger in workplaces can aid in improving workplace performance by solving interpersonal disputes [11]. The applications of anger detection also include personal anger management and aiding intelligent assistant systems such as Apple Siri or Amazon Alexa [7]. Recognizing anger in Interactive Voice Response (IVR) systems can be useful in designing dialog

strategies to correctly respond to the user by detecting his emotional state [12]. Therefore, anger detection from spoken conversations is receiving increasing attention in the speech research community. Besides, Bengali is the fifth most spoken native language and the sixth most spoken language in the world [13]. Despite being such a widely used language, there has been limited work on speech emotion recognition in Bengali [14, 15]. The existing works mostly focus on detecting emotion from pre-collected audio data. Analyses on audio-textual online emotion or anger recognition are not yet present in the Bengali language.

1.2 Background

Anger can be expressed through different mediums, such as audios, videos, texts, and physiological responses. Depending on the type of input data, anger detection can be classified into two types: traditional or offline anger detection, and early or online anger detection. Offline anger detection can be defined as the recognition of anger from pre-collected videos, audio, and texts, whereas online anger detection recognizes the anger of future timestamps from the present and/or past timestamps of videos, audios, or texts.

Anger can be conveyed using tone of speech and/or word choice [12]. In the context of spoken conversations, anger can be further classified into hot anger or cold anger [16], depending on how anger is expressed. While hot anger is expressed by changes in tone of speech that are indicative of anger such as yelling or shouting, cold anger is demonstrated with a flat or severe tone of speech accompanied by negative or threatening words. Therefore, though hot anger can be accurately detected using only acoustic features, cold anger detection requires textual features. Since cold anger is very common, especially in professional situations, considering linguistic features is extremely significant. Although textual transcriptions require speech-to-text conversion from the recorded audio, the importance of textual features in detecting cold anger makes the conversion worthwhile.

In the context of spoken conversations, anger can be detected from both acted and spontaneous conversations [17]. Acted conversations are recorded with the help of trained actors, who express different emotions in studio conditions. However, spontaneous conversations represent real emotions felt by the speakers without any external manipulation. A good example of spontaneous conversation is a call-center dialog. In call-centers, a huge number of conversations between callers and agents are recorded every day, which can be analyzed to improve the service of call centers.

1.3 State of the Art

Throughout the years, offline anger detection has been explored in speech communication research for both audio and textual data. However, online anger detection has not been explored in the audio-textual domain.

1.3.1 Research in Offline Anger Detection

Offline anger detection has been explored in different languages including English, French, Spanish, Arabic, Turkish, etc. [2, 12]. Some of them are based on only acoustic features [3, 18], while some are based on both acoustic and linguistic features [6, 10]. Most of the existing works are based on feature engineering, e.g., Mel-frequency cepstral coefficients (MFCCs), Linear predictive coding (LPC), pitch, energy, etc [1, 3, 6, 10, 12, 18]. However, existing works use feature engineering instead of raw speech waveform, which decreases the robustness of the model and increases dependency on feature extraction methods [19]. While most of the related works focus on traditional machine learning techniques [1, 6, 10, 12, 18], some works such as Deng et al. [3] focus on deep learning. However, the works which consider deep learning techniques for anger detection are audio-only and can not detect cold anger. Moreover, while gender information has been shown to improve the efficiency of speech anger detection, its effects have not been explored in detail. Abu Shaqra et al. [1] improved audio-only emotion detection by incorporating gender information. Nevertheless, their work is audio-only and will fail to detect instances of cold anger.

1.3.2 Research in Online Anger Detection

While online anger detection has applications in various potential fields such as personal anger management and aiding intelligent assistant systems [20], it has largely remained an unexplored area of research. To the best of our knowledge, there has been no previous work on the early detection of audio-textual anger. While there have been a few works on emotion forecasting [20], their domain was audio-only or audio-visual data. A related work by Shahriar et al. [20] works on the detection of the emotion of future audio-visual data. Additionally, another work closely related to ours by Mongkolnavin et al. [21] predicts forthcoming anger from Thai conversations. Shahriar et al. [20] and Mongkolnavin et al. [21] both consider the emotional flow of a single speaker in a dyadic conversation, even though the performance of emotion recognition can be improved if the influence of the other speaker is also taken into account [22]. Moreover, the approach by Shahriar et al. [20] will not work in normal audio conversations where visual information is not always available, especially in the case of call-center conversations. In addition, both of the research studies omit textual information, which does not allow the detection of

cold anger [16]. Additionally, they only utilize handcrafted acoustic features even though raw waveform features have been shown to consistently increase the robustness of a model [23].

1.3.3 Research in the Bengali Language

There is no existing work on audio-textual anger detection in Bengali. However, there are some research works that detect emotion using audio-only datasets, such as the works of Sultana et al. [24], Hasan et al. [14], Devnath et al. [4], Mohanta et al. [15], etc. Nevertheless, these works do not consider textual features, raw speech waveform and gender information, which can greatly improve the performance of emotion detection [1, 16, 19]. Moreover, despite the various potential applications of online detection, the existing works only perform offline detection. Besides, the existing works mainly focus on acted conversations instead of spontaneous conversations and work with much smaller datasets.

1.4 Problem and Overview of Solution

In this work, we detect utterance-level offline and online anger from audio-textual data. We model the input data as variable-length utterances which represent speaking turns in a conversation. For offline anger detection, the input is a single utterance, U and the output is its corresponding label, L . For online anger detection, the input is a sequence of utterances and the output is the anger label of the subsequent utterance/(s). In the simplified version of online anger detection, the input is a sequence of i utterances, (U_1, U_2, \dots, U_i) and the output is the anger label of the $(i + 1)$ -th utterance, L_{i+1} . A brief overview of the offline and online anger detectors can be found in Figure 1.1.

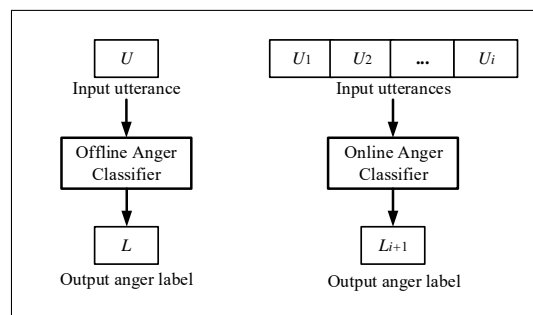


Figure 1.1: An overview of Offline and Online Anger Detectors

While anger detection has been explored previously, there exist a few challenges. The existing works rely on handcrafted features, whereas raw speech waveform features based on sinc filters have been shown to perform well for speech communication tasks [23]. Besides, most of the studies that use deep learning [3] do not incorporate textual information and will not work well

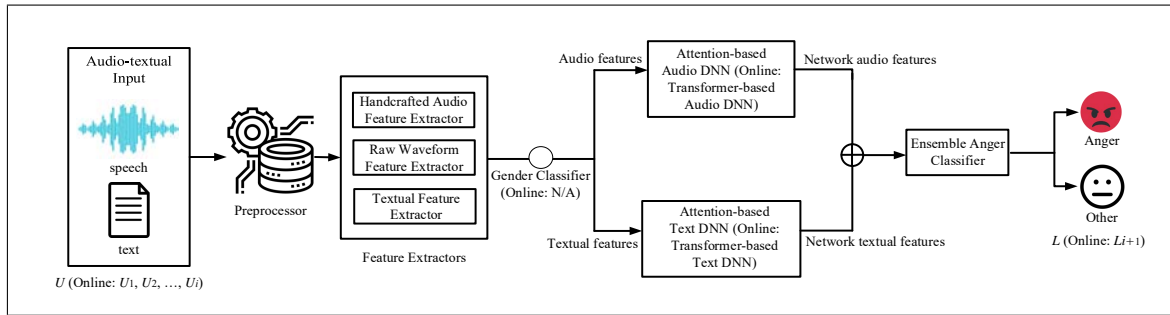


Figure 1.2: A detailed overview of the Offline Anger Detector, **OAD** and the Online Anger Detector, **EAD**

for detecting cold anger [16]. Moreover, the works on offline anger overlook the influence of gender information, even though gender information has been shown to improve the performance of audio-only emotion recognition [1]. Furthermore, to the best of our knowledge, there are no existing works on online audio-textual anger recognition. Related works are based on audio-visual or audio-only data, which do not work well for cold anger detection [20, 21]. Besides, they consider the utterances of a single speaker in a dyadic conversation despite the influence of the utterances of both speakers on the emotional flow of participants [22].

Therefore, we improve offline anger recognition by developing an ensemble-based deep CNN architecture for offline anger detection which incorporates gender information and self-attention mechanism [25]. Additionally, for online anger, we propose an ensemble-based transformer model which considers the utterances of both participants in a dyadic conversation. Both offline and online anger recognition approaches use textual features and sinc-based raw speech waveform features alongside handcrafted acoustic features.

A detailed overview of the Offline Anger Detector, **OAD** and the Online Anger Detector, **EAD** is presented in Figure 1.2. First, the input is pre-processed and a number of feature extractors are used to extract different handcrafted acoustic features and raw waveform features from the audio data, and textual features information from the transcribed audio. For offline anger detection, we incorporate gender information by classifying the gender of the speaker using a gender classifier and using separately trained anger classifiers for male and female speakers. Then, two neural network pipelines are used to extract deep audio and textual features. Finally, these features are fused in a mid-level fusion-based scheme and fed into the ensemble anger classifier which outputs an appropriate anger label. Online anger detection works similarly by incorporating audio and textual features in a mid-level fusion-based transformer architecture, which utilizes an ensemble downstream classifier.

Finally, we demonstrate the efficacy of our proposed approaches using two datasets: the Bengali call-center dataset and the IEMOCAP dataset. We prepare a transcribed version of our Bengali call-center dataset using Google speech-to-text [26] which contains approximately 33 hours of spontaneous conversational speech, and then annotate the data with anger labels. To evaluate our

models for offline and online anger, we compare the performances of our proposed models with several baseline methods including traditional machine learning models, such as Decision Tree, k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), etc., and deep learning models, such as Multilayer Perceptron (MLP), Long-Short Term Memory (LSTM) model, Transformer, etc. For online anger detection, we also vary different parameters such as the number of input utterances, number of skipped utterances, number of output utterances, etc., and observe the effects of the variation on the performance of the model. Furthermore, we demonstrate the effectiveness of our proposed models on the IEMOCAP dataset [27]. The IEMOCAP dataset is a widely used English dataset for emotion recognition that contains approximately 12 hours of audio and transcriptions of acted conversations. It is annotated into categorical emotion labels including anger and contains audio from both male and female participants.

1.5 Research Objective

In this study, we investigate the utility of audio-textual information to perform offline and online anger detection in dyadic conversations and develop effective machine learning techniques for anger detection. Thus, the primary objectives of our study are as follows:

- To perform offline audio-textual anger recognition by utilizing deep learning ensemble approaches while considering the impact of gender.
- To perform online audio-textual anger detection in dyadic conversations, considering the utterances of a single speaker and the utterances of both speakers.
- To analyze the efficacy of the proposed approaches on dyadic conversational speech datasets from Bengali and English.

1.6 Contribution

We summarize the key contributions of our research as follows.

- We develop an ensemble-based model to classify offline audio-textual anger by incorporating gender information, raw speech waveform, and deep learning techniques for improved performance. We also compare our proposed approach with different deep learning and traditional machine-learning algorithms. Moreover, we demonstrate that our model outperforms the baselines.
- To the best of our knowledge, we are the first to explore audio-textual online anger detection. We develop an ensemble-based transformer model to classify audio-textual

anger considering the influence of both speakers in dyadic conversations. We compare our model with different baseline models and demonstrate that our model consistently outperforms the others.

- We demonstrate the performance of our offline and online anger detection models on Bengali and English language datasets. We show that our approach performs better than the baseline feature sets in every case. Additionally, for online anger detection, we study the impact of varying the numbers of input, output, and skipped utterances on the performance of our model.

1.7 Organization

The rest of our thesis is organized as follows. Chapter 2 deals with the background information mentioned throughout our work. In Chapter 3, we discuss the previous studies relevant to offline and online audio-textual anger detection. In Chapter 4, we discuss offline anger detection in detail. In Chapter 5, we introduce and discuss online anger detection. Finally, we conclude the research in Chapter 6 and discuss our future works.

Chapter 2

Background

In this chapter, we present a detailed overview of different features used for audio-textual speech processing. Additionally, we discuss the relevant background information mentioned throughout this work, including different traditional machine learning models, deep learning models, and ensemble models.

2.1 Features used in Audio-Textual Speech Processing

Here, we discuss different audio and textual features used in the context of audio and textual speech processing. Commonly used audio features include loudness, pitch, intensity, etc. and textual features include Bag-of-Words (BoW), Term Frequency (TF), etc.

2.1.1 Audio Features

A broad variety of prosodic and acoustic audio features are used in speech signal processing. These features represent important information about vocal expression patterns [12]. The most commonly used features are defined as follows.

Loudness: Loudness is the amount of sound pressure in an audio signal [28].

Pitch: The pitch of an audio signal is its position in the complete range of sound [29].

Intensity: The intensity of sound is the power per unit area carried by a sound wave [29].

Zero-Crossing Rate: The rate at which signals change from negative to zero to positive or vice versa is known as the zero-crossing rate [30].

Fundamental Frequency: The fundamental frequency is the lowest frequency among the whole range of frequencies in an audio signal [31].

Formant Frequencies: Formant frequencies are the frequency peaks in the whole spectrum which correspond to high amounts of energy [29].

CHROMA Features: CHROMA features include 12 features that indicate how much energy of each of the 12 pitch classes is present in the audio [32]. It is highly related to the 12 different pitch classes of Western music.

CHROMA Energy Normalized Statistics (CENS) Features: CHROMA Energy Normalized Statistics (CENS) features take CHROMA feature statistics over large windows to smooth out fluctuations in tempo, articulation, etc [32].

Mel-Frequency Cepstral Coefficients (MFCCs): Mel-Frequency Cepstral Coefficients (MFCCs) are derived from the cepstral representation of a sound signal by mapping the frequencies to the mel scale so that the frequency bands are equally spaced [33].

Linear Predictive Coefficients (LPCs): Linear predictive coefficients (LPCs) are the compressed form of the spectral envelope of a digital audio signal [34].

Linear Predictive Cepstral Coefficients (LPCCs): Linear prediction cepstral coefficients (LPCC) are calculated from the spectral envelope of LPC [34].

2.1.2 Textual Features

Different textual features aid in audio-textual speech emotion processing, including Bag-of-Words (BoW), Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), etc. These features are discussed as follows.

Bag-of-Words (BoW): The bag-of-words model is a model in which a text is represented using a collection of words [12]. In this model, grammar, ordering of words in a sentence, etc. information is discarded. It is commonly used for document classification.

Term Frequency (TF): Term frequency indicates the frequency of a particular word in a document [12]. For a particular word w in document d , the term frequency is,

$$tf(w, d) = \text{count of } w \text{ in } d / \text{number of words in } d$$

Term Frequency-Inverse Document Frequency (TF-IDF): Term Frequency-Inverse Document Frequency (TF-IDF) indicates the importance of a particular word in a document among a collection of documents [12]. If there are D documents in a data set, where the document frequency of word w is df , and the term frequency of w in document d is $tf(w, d)$, then TF-IDF of w in d is calculated as follows.

$$tf - idf(w, d) = tf(w, d) \times \log(|D| / (df + 1))$$

Self-Referential Information (SRI): Self-Referential Information (SRI) represents the information of a word with respect to an emotion class [12]. If $w \in W$ is a word in a vocabulary, $e \in E$ is an emotion class, $P(e)$ is the prior probability of an emotion class, and $P(e|w)$ is the

posterior probability that a certain word implies a certain emotion, then SRI can be expressed as follows.

$$SRI(e, w) = \log \frac{P(e|w)}{P(e)}$$

Conversational Emotional Salience: Conversational Emotional Salient words (CEMS) are obtained from a corpus by Pearson’s chi-square test which is used to determine word-frequency differences [16]. For each emotion class and speaker, Pearson’s chi-square test is conducted. Additionally, word count is computed for the words in all utterances in the data set, from which the expectation of word count is computed. Then, the chi-square score is calculated from the word count and expectation of word count. The words that are considerably different in terms of frequency are extracted as CEMS.

2.2 Models and Techniques used in Audio-Textual Speech Processing

In this section, we discuss various models and techniques which will be mentioned throughout this work, including traditional machine learning models, deep learning models, and ensemble models.

2.2.1 Traditional Machine Learning Models

Throughout the years, traditional machine learning techniques have been used in different domains, including anger recognition. Here, we discuss different traditional machine learning models that were used in the previous works on anger recognition, such as decision trees, k-nearest neighbours (k-NNs), and support vector machines (SVMs).

Decision Trees

Decision Trees learn simple decision rules from the training data to predict the class or value of the input variable [35]. In decision trees, each node performs a test on a feature and passes the computation to the next node following the branch corresponding to the output of the test. Finally, the output is obtained when the computation reaches the leaf nodes. It is one of the most commonly used baseline models in machine learning tasks.

k-Nearest Neighbour (k-NN)

k-Nearest Neighbour (k-NN) is one of the simplest machine learning algorithms which classifies a target data by computing its similarity with its neighbours by using an appropriate algorithm [36].

The target data is assigned the most common class among its k neighbours by majority voting.

Support Vector Machine (SVM)

Support Vector Machine (SVM) models map the input data to a high-dimensional feature space. Then, the data points can be classified even when they are not linearly separable [37]. SVMs with polynomial kernel functions allow learning of non-linear models by representing the similarity of vectors (training samples) in a feature space over polynomials of the original variables [38].

2.2.2 Deep Learning Models

We discuss different deep learning models for online and offline anger detection. These models include fully-connected deep neural network (DNN), recurrent neural network (RNN), long short-term memory network (LSTM), bidirectional long short-term memory network (Bi-LSTM), convolutional neural network (CNN), transformer, etc.

Fully-Connected Deep Neural Network (DNN)

Fully-connected deep neural networks consist of fully connected or dense layers, which are feed-forward neural networks. In dense layers, each node of a layer is connected with a particular weight to every node in the next layer [39]. As a result, these layers allow the learning of features from all combinations of the previous layer. Each node in a dense layer uses an activation function.

Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is a type of Artificial Neural Network (ANN) [40] that has feedback connections. It uses tanh activation function and contains no memory units or gates.

Long Short-Term Memory Network (LSTM)

Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) [41]. These networks use tanh function as the activation function. LSTMs are different from simple RNNs due to the presence of gates: forget gate, input gate, and output gate. These gates control the flow of information through the structure which allows the model to have longer-term dependencies, making LSTMs superior to simple RNNs.

Bidirectional Long Short-Term Memory Network (LSTM)

Bidirectional Long Short-Term Memory Networks (Bi-LSTMs) consist of two LSTMs as shown in Figure 2.1. The first LSTM processes the input in a forward direction, while the second LSTM

processes the input in a backward direction [42]. The outputs of the forward and backward LSTMs are then fed into the activation layer. The output of the activation layer is considered the final output of the Bi-LSTM. In most cases, Bi-LSTMs perform better than LSTMs due to traversing the input data twice.

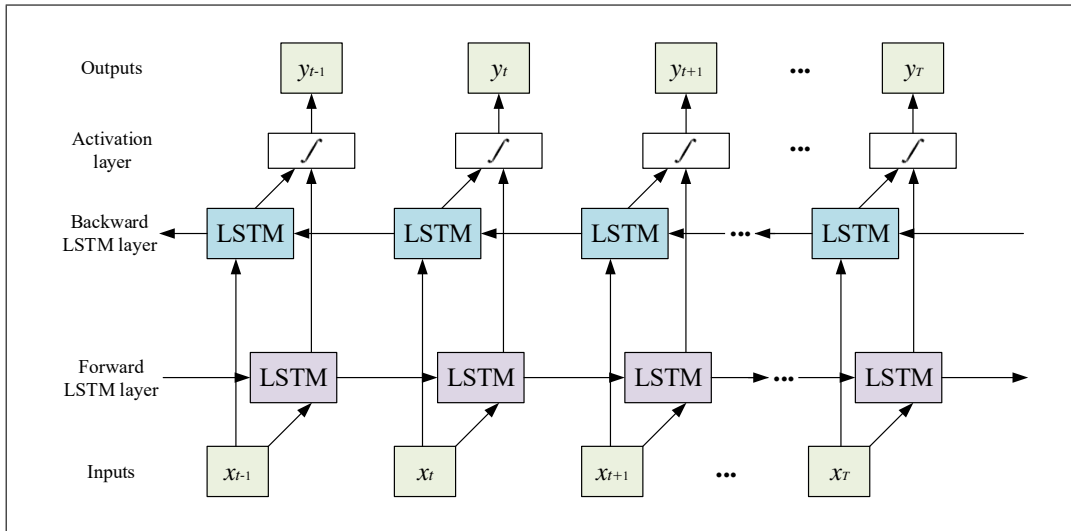


Figure 2.1: The architecture of a Bidirectional-LSTM (BiLSTM) network

Convolutional Neural Networks (CNNs)

Convolutional neural networks are feed-forward Artificial Neural Networks (CNNs). Convolution layers convert input data into an abstract feature map [43] and can extract spatial features from the data using kernels. Since 1-D CNNs are computationally simple, their implementation is less costly. In this work, we use 1-D CNNs containing max pooling layers, global average pooling layers, and batch normalization layers. These layers are discussed as follows.

1-D Convolution Layer: In Conv1D or 1-dimensional convolution layers, the kernel slides along one direction only [44]. This layer consists of n kernels or filters and performs a convolution operation between the input vector and the filter. Then, it produces an output vector that has n channels. Since the output of convolution layers is smaller than the input, the final output of the network becomes very small if multiple such layers are stacked together. Therefore, padding is used to increase the size of the input, so that the output size is equal to the original input size. In this work, we have used zero padding and ReLU activation.

Batch Normalization Layer: Batch normalization normalizes the output of neural network layers and is used to stabilize neural networks [45]. Batch normalization reduces internal covariate shift and allows faster training alongside producing more reliable models. It also allows the use of higher learning rates without the problem of vanishing gradients. Besides, it eliminates the need for Dropout by acting as a regularizer. By using batch normalization, the network becomes more robust regarding different initialization mechanisms.

Max Pooling Layer: Convolutional neural networks include pooling layers that reduce the dimensionality of the data. They are usually used after one or more convolution layers. They combine the outputs of multiple neurons in each section of each feature map. In the maximum or max pooling layer, only the maximum value among a cluster of neurons of the feature map is considered [46]. Max pooling reduces data invariances and downscales the data by extracting the most significant features.

Global Average Pooling Layer: Global average pooling is a type of pooling mechanism used in CNNs. The global average pooling layer generates a feature map for each corresponding category of the classification task in the last convolution layer by taking the average of each feature map and is generally used instead of Flatten layers [47]. Moreover, using this layer avoids the overfitting problem since there are no parameters to optimize.

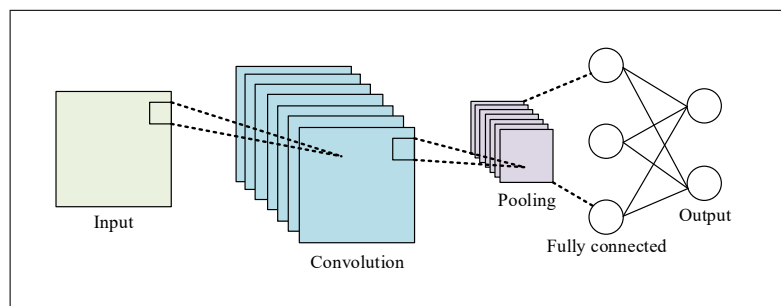


Figure 2.2: The architecture of a Convolutional Neural Network (CNN)

An example of a CNN with two output classes is shown in Figure 2.2. The network consists of an input layer, one convolution layer, one pooling layer, one fully-connected or dense layer, and an output layer.

Transformers:

A Transformer is a deep learning network that uses self-attention in an encoder-decoder framework [48]. Attention highlights some portions of the input data while diminishing others, similar to how humans pay more attention to more relevant aspects of information while paying less attention to others. Transformers are good at handling sequential input data, particularly in the context of natural language processing.

The architecture of the Transformer proposed by Vaswani et al. [48] is shown in Figure 2.3. The Transformer employs the encoder-decoder architecture, with stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, as shown in the left and right halves of Figure 2.3.

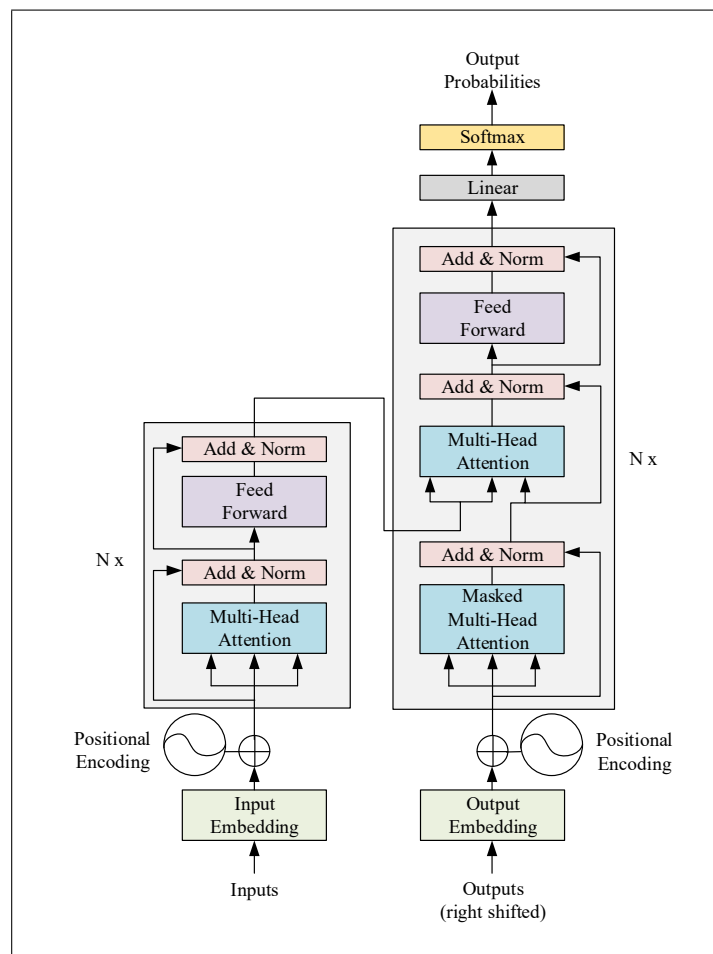


Figure 2.3: The architecture of a Transformer network

2.2.3 Ensemble Machine Learning Models

This work mentions different ensemble machine learning models as well. Here, we discuss random forests, AdaBoost, Gradient Boost, and XGBoost.

Random Forests

Random Forest is an ensemble of multiple decision trees constructed using feature bagging. The predictions from all the decision trees are considered, yielding better results than any single tree [49]. Random forests are used for both classification and regression tasks. For classification and regression tasks, random forest considers the majority voting of the trees and the average prediction of the trees respectively.

AdaBoost

Adaptive Boosting or AdaBoost is an ensemble classifier consisting of several weak learners in which the following weak learners are tuned to favor instances wrongly classified by the previous

classifiers [50]. Here, the weak learners are decision trees with a single split. It creates a model by assigning equal weights to all data points. Then, it assigns higher weights to incorrectly classified points. In the following model, all points with higher weights are considered more important. The algorithm will continue to train models until a lower error is received. AdaBoost tries to minimize exponential loss.

Gradient Boost

Gradient Boost is an ensemble classifier consisting of weak prediction models or decision trees [51]. It is an additive model where the existing trees are not changed, rather new decision trees are sequentially added one at a time. Gradient descent optimizes training loss using the partial derivative of the loss function. It also allows the optimization of any differentiable loss function. Since any differential loss function can be used for Gradient Boost, it is more robust compared to AdaBoost.

XGBoost

XGBoost or Extreme Gradient Boosting is a scalable ensemble classifier [52]. The Gradient Boost in XGBoost comes from the work of Friedman [51]. Compared to Gradient Boost, XGBoost is more regularized and reduces overfitting even with less training data. XGBoost minimizes an objective function which is a regularised differentiable convex loss function measuring the difference between the prediction and the target. It uses a second-order partial derivative of the loss function which provides a better approximation compared to gradient boosting. It also implements $L1$ and $L2$ regularization and parallel computing. Besides, XGBoost can handle sparse datasets and datasets with missing values.

Chapter 3

Related Work

The related work chapter of our study consists of the discussions on speech communication along with offline and online emotion recognition considering different techniques and feature sets. We also explore the existing related works in the Bengali language.

3.1 Research in Speech Communication

There are many research studies in the field of speech communication, notably on speech recognition and speech synthesis. Speech Emotion Recognition (SER) is a particular branch of speech communication research that has been explored by various research studies [53, 54]. Anger recognition, in particular, has also been studied in the context of conversational speech [17]. Besides, other branches of speech communication research include speech recognition [55], speaker recognition [56], and speech synthesis [57].

3.2 Research in Emotion Recognition

Emotion recognition has been an active research area in the field of speech communication. Research studies commonly perform emotion detection from audio features, textual features, or audio-textual features [58]. While some of them use traditional machine learning models, others use deep learning algorithms [59].

3.2.1 Emotion Recognition from Speech

Emotion recognition from speech has gained considerable attention over the last several years. Consequently, emotion recognition from audio features has been attempted in many languages including English [60, 61], German [62], Chinese [63], Turkish [64], Russian [65], and many other languages. Wang and Guan [54] detect emotion from speech using prosodic and spectral

features. They employ different classification techniques including the Gaussian mixture model (GMM), Neural network (NN), K-nearest neighbors (K-NN), and Fisher's linear discriminant analysis (FLDA). Besides, Yeh et al. [66] perform emotion recognition from speech using deep learning techniques with Gated Recurrent Units (GRUs). Parry et al. [67] also recognize emotion from spoken conversations with the help of deep learning algorithms such as Convolutional neural networks (CNNs) and Long-short term memory networks (LSTMs). Zaman et al. [68] detect emotion, gender, and age from speech. They use different models for detection, such as Random Forest, CatBoost, Gradient Boosting, K-nearest neighbors (KNN), XGBoost, AdaBoost, etc. They show that, for emotion detection from speech, XGBoost performs better than the other models.

3.2.2 Emotion Recognition from Text

Text emotion recognition or sentiment analysis is a much-explored domain in the field of emotion recognition. Wu et al. [69] propose a method to recognize emotion from textual data with the help of a Separable mixture model (SMM). Shaheen et al. [70] utilize automatically generated emotion recognition rules (ERRs) to perform emotion recognition. They use classification algorithms such as K-nearest neighbors (KNNs), Point Mutual Information (PMI), and PMI with Information Retrieval (PMI-IR). Phan et al. [71] also perform textual emotion detection. However, they utilize conversational transcripts from a movie dialog corpus.

3.2.3 Multimodal Emotion Recognition

Several existing works recognize emotion from multimodal datasets [72]. Huang et al. [73] propose a novel Hierarchical LSTM architecture to recognize emotion from audio-textual conversations. Cai et al. [74] also perform emotion detection from audio-textual data by combining CNNs, LSTMs, and Bi-LSTMs in a novel architecture. Besides, Hazarika et al. [75] propose an Interactive COnversational memory Network (ICON) which extracts various features from video data and classifies them into emotion classes. During classification, they consider self as well as inter-speaker emotional influences.

3.3 Research in Offline Anger Recognition

Offline or traditional anger recognition has been studied across different mediums such as audio, text, video, physiological changes of the body, etc [76–80]. Offline anger recognition has also been explored across different languages, feature sets, modalities, and techniques.

3.3.1 Offline Anger Recognition across Different Languages

Anger recognition has been performed in various languages including English, Chinese, German, Swedish, Arabic, Turkish, etc. Neiberg et al. [17] explore anger recognition from spontaneous Swedish speech using telephone speech obtained from the Swedish company Voice Provider which provides voice-controlled telephone services. Polzehl et al. [12] perform anger detection from acoustic and prosodic features obtained using 10 hours of American English and 21 hours of German telephone conversations. Pohjalainen et al. also recognize anger from German telephone speech [81]. In addition, Khalil et al. [6] perform anger detection from Arabic speech dialogs using an Arabic speech corpora constructed from real-world dialogs. Moreover, Erden et al. [10] study the recognition of anger in Turkish call-center conversations. They include textual features by manually transcribing all conversations. Besides, ElShaer et al. [82] perform anger recognition from Chinese speech corpora. Mongkolnavin et al. detect anger in Thai call-center conversations [21]. They consider 100 conversations from a real Thai call-center and use Good Guess, Random Guess, and LSTMs to detect anger. They find that LSTMs perform better than the baseline techniques.

3.3.2 Offline Anger Recognition across Different Modalities and Feature Sets

Different models have been developed for anger detection across different modalities, which use different feature sets and classifiers. Some works on anger detection deal with audio features only. Khalil et al. [6] perform anger recognition using a Support Vector Machine (SVM) model using acoustic features such as fundamental frequency, formants, energy, and Mel-frequency cepstral coefficients (MFCCs). Pappas et al. [18] describe a method to classify anger from speech using features such as energy, pitch, MFCC, etc. using a Logistic Regression (LR) classifier. Neiberg et al. [17] explore anger recognition from speech using Linear Discriminant Analysis (LDA) and Gaussian Mixture Model (GMM) classifiers. Lee et al. [83] detect high and low anger in speech data using a Back-Propagation Network (BPN) with acoustic features. Besides, ElShaer et al. [82] recognize anger from speech using transfer learning. They employ SoundNet to detect anger, which is a Convolutional Neural Network (CNN) model trained on a huge video dataset. Deng et al. [3] also use deep learning models such as Deep Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs), Convolution Neural Networks (CNNs), etc. to detect anger from speech data.

Moreover, some relevant studies work on detecting anger from both audio and textual features. Polzehl et al. [12] perform anger recognition from both acoustic and linguistic features. For acoustic modeling, they use pitch, loudness, MFCC, etc. For linguistic modeling, they use Bag-of-Words (BOW), Term Frequency (TF), Term Frequency - Inverse Document Frequency (TF.IDF), and the Self-Referential Information (SRI). Nomoto et al. [16] also perform anger

recognition from audio-textual conversations. They classify the utterances into Hot Anger, Cold Anger, and Neutral using an SVM model. The linguistic features used in this work represent conversational emotional salience. They show that considering both acoustic and textual features perform better compared to considering only acoustic features.

3.3.3 Offline Anger Recognition Incorporating Different Techniques

Some research studies indicate that the performance of anger recognition can be improved by incorporating various techniques. Abu Shaqra et al. [1] show that emotion recognition, including anger, can be improved by including gender information. They use two separate classifiers for male and female speakers. First, they classify the gender of the speaker using a gender classifier. Then, this information is used to decide whether to use the male classifier or the female classifier. They use the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for gender, age, and emotion recognition. Their findings revealed that using a separate emotion model for each gender and age category yields higher accuracy than using a single classifier for all data. Priyasad et al. [19] demonstrate that the performance of anger recognition can be improved by using the raw audio waveform. Additionally, a study by Zheng et al. [84] shows that emotions, including anger, can be classified more efficiently if ensemble machine learning models are used.

3.4 Research in Online Anger Recognition

Even though online recognition or forecasting has received considerable attention from different domains [85–87] to the best of our knowledge, there is no existing work on online audio-textual anger recognition. However, Kim et al. [88] work on audio-visual data and demonstrate that emotion recognition using only a portion of data from an utterance achieves accuracy comparable to using the full data within an utterance. This indicates that the effect of emotion can be observed even from partial utterances. Besides, Noroozi et al. [89] perform emotion forecasting from speech using manually created time series by concatenating isolated emotional speech. Nevertheless, their work deals with manually concatenated utterances instead of spontaneous dyadic conversations.

Moreover, a related work to ours by Shahriar et al. [20] deals with predicting emotion from audio-visual data using Fully-Connected Deep Neural Networks (FC-DNN), Deep Long Short-Term Memory (D-LSTM), and Deep Bidirectional Long Short-Term Memory (D-BLSTM) Recurrent Neural Networks (RNNs). They employ past audio-visual cues alongside the present utterance to forecast future emotions from the utterances of a single speaker. They demonstrate that history-added forecasting performs better compared to considering the present utterance only. They compare the performance of forecasting using the previous utterance with a randomly added utterance and show that the previous utterance performs better. They also show that

D-BLSTM and D-LSTM outperform FC-DNN for emotion forecasting and demonstrate their performances on the IEMOCAP [27] dataset. Besides, another closely related work to ours by Mongkolnavin et al. [21] predicts forthcoming anger from Thai call-center conversations. They utilize LSTMs, Good Guess, and Random Guess to predict anger from the utterances of the customers only and find that LSTMs perform better compared to the other techniques.

The research studies by Shahriar et al. [20] and Mongkolnavin et al. [21] both consider the utterances of a single speaker. However, emotion recognition performs better when the utterances of the other speaker are considered in a dyadic conversation since the emotions of one speaker influence the other [22]. In addition, the work by Shahriar et al. [20] requires visual information, whereas visual information is not always available, especially for call-center or telephone recordings. Furthermore, both of the studies fail to detect cold anger, expressed with a severe or flat tone of speech alongside harsh words, since they use either audio or audio-visual features. To detect cold anger, textual features are required. Additionally, while extracting audio features, both works consider only handcrafted acoustic features, whereas raw speech waveform features have been shown to increase the performance and robustness of models [23].

3.5 Research in the Bengali Language

Despite being such a widely spoken language, there are no existing works on audio-textual anger recognition in the Bengali language. However, there exist a few works which deal with emotion recognition in Bengali. Some of the existing works deal with emotion recognition from Bengali text corpora, while others deal with emotion recognition from Bengali speech corpora.

3.5.1 Emotion Recognition from Bengali Speech

There are some works in Bengali that deal with emotion recognition from speech. Sultana et al. [24] detect emotion from audio-only Bengali speech dataset SUBESCO using Deep CNN and Bi-LSTM networks. Additionally, they perform cross-corpus training, multi-corpus training, and transfer learning was employed for the Bangla and English languages using the SUBESCO and RAVDESS datasets. Mohanta et al. [15] detect emotional states using a Support Vector Machine (SVM) model on a self-built acted Bengali speech corpus. They use audio feature vectors such as Linear Prediction Cepstral coefficient (LPCC), Mel-frequency Cepstral Coefficient (MFCC), pitch, intensity, and formant for the detection. Devnath et al. [4] perform emotion recognition by using a self-built acted and isolated Bengali speech corpus. They use a k-Nearest Neighbor classifier with audio features such as pitch and Mel-frequency Cepstral Coefficient (MFCC). Similarly, Hasan et al. [14] use Mel-frequency Cepstrum Coefficient (MFCC) and Modulation Spectral (MS) features to recognize emotion from a self-built acted Bengali speech dataset. They employ a Recurrent Neural Network (RNN) model for classification purposes.

However, the existing works on speech emotion recognition do not use textual features which can help in the detection of cold anger [16]. Besides, the existing works do not use gender information and speech waveform, which can improve the performance of anger detection and emotion recognition [1, 19]. Additionally, the existing works are on much smaller datasets and do not consider spontaneous speech, which is significantly different from acted speech and comes with additional challenges [2].

3.5.2 Emotion Recognition from Bengali Text

The existing works on emotion recognition from Bengali text corpora deal with both article-level and sentence-level sentiment analysis. Tuhin et al. [90] use two machine learning approaches: the Naive Bayes Classification Algorithm and the Topical approach for performing sentiment analysis. By performing comparative analysis, they show that the Topical approach performs better for both article-level and sentence-level sentiment analysis. Tripto et al. [91] detect sentiment from Bengali text data obtained from YouTube comments. They employ deep learning-based models to classify emotion based on a three-class (positive, negative, neutral) and a five-class (strongly positive, positive, neutral, negative, strongly negative) sentiment label. They evaluate the performance of their model on Bengali, English, and Romanized Bengali comments from YouTube videos. Rafi-Ur-Rashid et al. [92] perform Bengali text classification focusing on minority classes in cases of class imbalance in datasets. They perform sentiment analysis with the help of BiLSTM, CNN, BiGRU, and LSTM models. They also incorporate ensembling with the help of these deep learning models. Besides, they use multiple linear and tree models which are trained using the hidden feature space of a BiLSTM model. Azmin et al. [93] detect emotion from Bengali text data collected from Facebook comments on different topics. For classification purposes, they employ a Naive Bayes (NB) classifier. Their feature set includes stemmer, Parts-of-Speech (POS) tagger, N-grams, and Term Frequency-Inverse Document Frequency (tf-idf). Asimuzzaman et al. [94] perform sentiment analysis using Bengali microblogs. They utilize an Adaptive Neuro-Fuzzy Inference System for the task in order to consider the subjective nature of the sentiments without assigning an absolute polarity. Taher et al. [95] perform opinion mining on diverse web-based Bengali text data obtained from different social media sites. They utilize the N-gram method for vectorization along with both linear and nonlinear Support Vector Machine (SVM) models for classification. They show that the performance of the N-gram is better when a particular opinion is expressed by two or more sentences.

Chapter 4

Offline Anger Detection

In this chapter, we propose an anger detection mechanism from pre-existing audio-textual data, which will be termed offline anger detection. In this model, we include a gender detection mechanism to improve the performance of anger detection.

4.1 Methodologies

Firstly, we describe the insights behind our approach. Then, we present an overview of our gender classifier. Lastly, we discuss our offline anger detection model in detail, which is termed **Offline Anger Detector (OAD)**.

4.1.1 Intuitions behind the Proposed Approach

The existing works on offline anger detection mostly use traditional machine learning techniques. Moreover, the existing works overlook the influence of gender information on anger, even though gender information has been shown to improve the performance of audio-only emotion recognition [1]. Additionally, the existing works rely on handcrafted acoustic features, whereas raw speech waveform features based on sinc filters have been shown to perform well for speech communication tasks [23]. Furthermore, most studies on anger detection do not incorporate textual information. As a result, their approach will not work well for detecting cold anger [16]. Therefore, we improve offline anger recognition by developing an ensemble-based deep learning approach for offline anger detection which incorporates textual features, gender information, and sinc-based raw speech waveform features alongside handcrafted acoustic features. While raw speech waveform features work well on larger datasets, handcrafted acoustic features perform better for smaller sample sizes [23]. Therefore, we consider both handcrafted acoustic features and raw speech waveform features in order to tackle datasets of various sizes.

4.1.2 Feature Extraction

For offline anger detection, we use both audio and text features. For audio, we use handcrafted acoustic features obtained from openSMILE emobase2010 [96] and raw waveform features from SincNet [23]. For textual features, we use BERT base [97] for the IEMOCAP dataset and BanglaBERT [98] for the Bengali call-center dataset.

4.1.3 Feature Extraction from Audio

During audio feature extraction, we model the data as utterances. We use Voice Activity Detection (VAD) to determine the start and end time of an utterance. Therefore, the utterances can be of variable length. Besides, we eliminate the noisy segments where both speakers are speaking together. By detecting simultaneous activity in both channels, such noisy segments can be eliminated.

In this work, we use both raw waveform features and handcrafted acoustic features. Handcrafted acoustic features perform well for smaller sample sizes. Besides, these features are less ambiguous and have increased interpretability. However, raw waveform features reduce dependency on the feature extraction methods and work well with large volumes of training data. For our classifier to perform well for both large and small datasets, we include raw waveform features alongside handcrafted acoustic features.

Raw Waveform Features: SincNet

We employ SincNet [23] to extract raw waveform features from the audio files. SincNet is a novel Convolution Neural Network (CNN) architecture, based on parameterized sinc functions. The sinc functions implement rectangular band-pass filters, where each rectangular band-pass filter can be thought of as two low-pass filters with learnable cutoff frequencies. This allows the architecture to extract more meaningful features by learning custom filter banks. For tasks such as emotion recognition and speaker recognition, SincNet has been shown to have higher interpretability and faster convergence rates compared to standard convolution networks [23]. It is also computationally efficient and requires only a few parameters. The convolution operation performed in SincNet can be expressed by the following equation.

$$y[n] = x[n] * g[n, \theta] \quad (4.1)$$

Here $x[n]$ is the n -th chunk of the input speech signal, $y[n]$ is the filtered output, g is the filter-bank function representing a filter-bank composed of rectangular bandpass filters, and θ represents the learnable parameters.

The function g can be represented in the time domain by inverse Fourier transform using the

following equation where f_1 is the low cutoff frequency, f_2 is the high cutoff frequency, and $\text{sinc}(x) = \sin(x)/x$.

$$g[n, f_1, f_2] = 2f_2 \text{sinc}(2\pi f_2 n) - 2f_1 \text{sinc}(2\pi f_1 n) \quad (4.2)$$

SincNet consists of multiple layers where the first layer performs sinc-based convolutions mentioned in equation 4.1. The first layer consists of 80 filters of size 251. The second and third layers are convolution layers of 60 filters of size 5. All the convolution layers and the input sample use batch normalization as a regularizer. The fourth, fifth, and sixth layers consist of fully connected layers with 2048 neurons each.

Handcrafted Audio Features: openSMILE

We use the openSMILE feature extraction toolkit [99], more specifically, the ‘emobase2010’ feature set for our work, which is based on the INTERSPEECH 2010 Paralinguistic Challenge feature set [96]. This feature set is specifically recommended for emotion recognition. OpenSMILE extracts audio low-level descriptor features from the input file, such as CHROMA, CHROMA energy normalized statistics (CENS), loudness, Mel-frequency cepstral coefficients (MFCCs), Linear predictive coefficients (LPCs), Line spectral frequencies, Fundamental frequency, Formant frequencies, etc. It also extracts delta regression and different statistical features such as mean, maximum, minimum, standard deviation, etc. from the low-level descriptors.

Emobase2010 Feature Set: This set consists of 1582 features. There are 34 low-level descriptors (LLDs) with their corresponding delta coefficients [96]. It also includes 21 functionals corresponding to each of the 68 LLD contours. Finally, it contains 19 functionals that are applied to the pitch-based LLDs and their corresponding delta coefficient contours along with the number of pitch onsets.

4.1.4 Feature Extraction from Text

Since we deal with conversational audio in this work, the audio is first separated into utterances which are then transcribed into text. To extract textual features from these transcriptions, we employ BERT [97] and BanglaBERT [98].

Feature Extraction from English Text: BERT

We use BERT base to extract features from the English text, which is a transformer-based model consisting of 12 encoder layers stacked together. It further includes 12 attention heads and 110 million parameters. The output size for each word is 768 dimensions. The model takes an input

of 512 dimensions and if the input size is larger than 512 words, we only use the first 512 words, as the original BERT. If the input size is smaller, then we perform padding at the end of the sentence and generate a mask of tokens for the input in which the indexes of the original words of the sentence contain '1', and the rest contain '0'.

Moreover, the input sentence is tokenized using a tokenizer into a token embedding. [CLS] token is added at the beginning of the token embedding and [SEP] token is added at the end. If the input consists of multiple sentences, then [SEP] token is used between each sentence. Besides, a segment embedding is also required in which the indexes of the words of the first sentence contain '0', words of the second sentence contain '1', and so on. Then, the input is fed into the model, which generates an output for each input token. The output of the model is the hidden state of the last layer of the model. For each token, BERT base generates an output of size 768, which is the predefined size of the last hidden layer.

Feature Extraction from Bengali Text: BanglaBERT

We use BanglaBERT [98] in order to extract features from the Bengali text. This model is inspired by the ELECTRA model [100], which is essentially a pre-training approach combined with BERT. This model jointly trains a generator and a discriminator. Some masked tokens are used as input to the generator which then predicts the original tokens for all masked tokens. Instead of using masked tokens as the discriminator input, tokens sampled from the output distribution of the generator for those masks are used. The discriminator then predicts whether each token is from the original input or replaced by the generator. The generator is discarded after pre-training, keeping only the discriminator model for downstream tasks. ELECTRA's model architecture and most hyperparameters are the same as BERT's. BanglaBERT uses ELECTRA-base, which is mostly equivalent to BERT-base.

4.1.5 Offline Anger Detection

For offline anger detection, the audio data is pre-processed, which is then used to classify whether the speaker is male or female. Afterwards, we detect anger from both audio and text by using separately trained anger classifiers for male and female speakers. The overall model is depicted in Figure 4.1.

Gender Detection Model

During gender detection, we only consider the audio data, since the transcriptions of utterances are usually not required for gender detection [101]. We adopt the model developed by Alkhaldeh et al. using Convolutional Neural Network (CNN) classifiers [101]. However, we

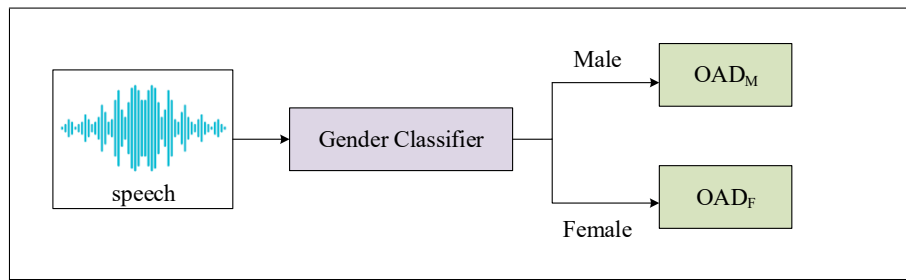


Figure 4.1: Traditional or Offline Anger Detector with Gender Classifier

make it more efficient by using openSMILE for handcrafted feature extraction and incorporating raw speech waveform features in a mid-level fusion architecture.

At first, the conversational audio is separated into utterances. Each utterance is further separated into $250ms$ chunks. For each chunk, we obtain the raw waveform features and the handcrafted features from the pre-processed data, which are then fed into two separately trained deep-learning CNN classifiers. Then, we combine the outputs from the two classifiers and use a mid-level fusion scheme to classify the gender of the speaker using a deeply connected layer along with the softmax output of the final layer. Finally, we obtain the response for the whole utterance by combining the responses obtained from each chunk. The gender classifier is shown in Figure 4.2. The different blocks used in the gender classifier are discussed in Subsection 2.2.2.

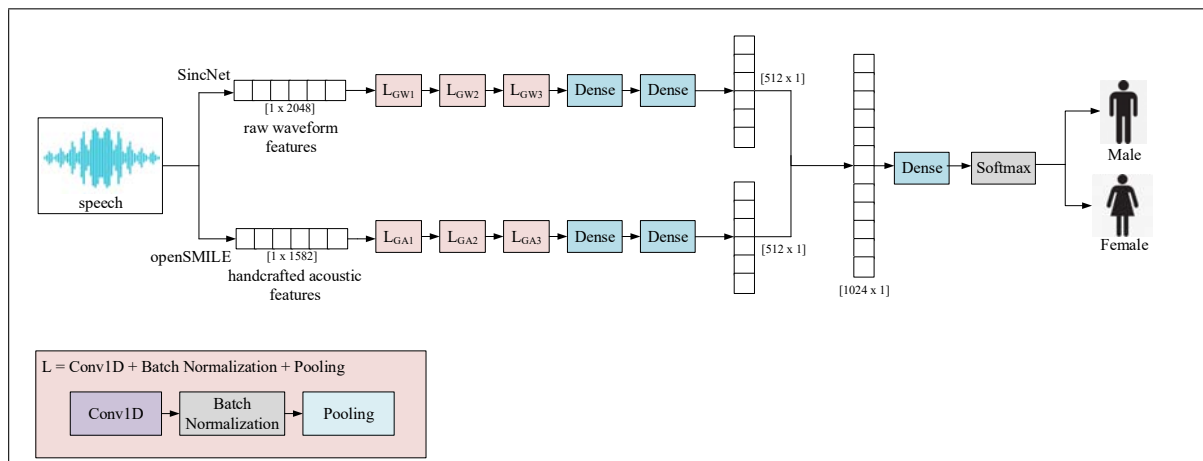


Figure 4.2: Gender Classifier

Now, we present a complete description of our gender classification model. For each audio chunk obtained from an utterance, we extract the raw waveform features of dimension $[1 \times 2048]$ using SincNet. We additionally extract some important handcrafted features of dimension $[1 \times 1582]$ using openSMILE. We use two separate pipelines for these features. Each pipeline is a CNN consisting of three sets of Convolution 1D, Batch Normalization, and Pooling layers. Each of the convolution layers has ReLU activation. The first convolution layer consists of 32 filters of size $= 3$ and stride $= 1$. The second convolution layer has 64 filters. Each of those filters is of size $= 3$ and stride $= 1$. The third convolution layer has 128 filters with filter size $= 3$ and stride $= 1$.

All the pooling layers apart from the final pooling layer have filter size = 2 with stride = 2 and perform max pooling. The final pooling layer performs global average pooling. We show the convolution1D, batch normalization, and pooling layers as a single block termed L in Figure 4.2. In addition, for each CNN pipeline, the output of the final pooling layer is fed into two subsequent fully connected (FC) or dense layers consisting of 512 nodes each with ReLU activation. Then, the two output vectors of the CNN pipelines of size $[512 \times 1]$ are concatenated to form a vector of size $[1024 \times 1]$. This acts as the input of a final dense layer and a softmax layer which outputs the probabilities of the speaker being male or female. Then, the classifier outputs the gender with the maximum probability indicated by the softmax layer.

Offline Anger Detector (OAD)

We use separately trained offline anger classifiers for male and female speakers. Since the structures of the models for male and female speakers are the same, we will refer to them as Offline Anger Detectors (OAD) from here on. Most of the blocks used in the model, such as convolution 1D, batch normalization, max pooling, etc. are described in Subsection 2.2.2. Additionally, our offline anger detection model contains self attention mechanism, which is explained as follows.

Self Attention Mechanism

Attention maps a query and a set of key-value pairs to an output, which is the weighted sum of the values [48]. The weight assigned to each value is obtained using a function of the query with the associated key. The function of attention is to mimic cognitive attention. Similar to how humans pay more attention to more relevant parts of information while paying less attention to the others, attention highlights some portion of the input data and diminishes the others [48]. We calculate self attention using the work of Fernando et al. [25]. If the input vector is i_t , the neural network output vector is k_t , the attention score is α_t , the context vector is \hat{k} , and the output of the attention layer is o , then self attention can be computed using the following equations.

$$k_t = \tanh(\mathbf{W}_k \mathbf{i}_t + a_h) \quad (4.3)$$

$$\alpha_t = \frac{\exp(|\mathbf{k}_t|^\top \hat{\mathbf{k}})}{\sum_t \exp(|\mathbf{k}_t|^\top \hat{\mathbf{k}})} \quad (4.4)$$

$$o = \sum_t \alpha_t \times \mathbf{i}_t \quad (4.5)$$

Offline Anger Detector (OAD): The Complete Model

Now, we present a complete description of our offline anger detector model. Similar to the gender classifier, the conversational audio is split into utterances, which are further split into 250ms segments. For each segment, we obtain the response of the classifier, and the response for the whole utterance is found by adding the responses obtained from each segment to get the final classification scores, similar to SincNet [23].

For each chunk, we obtain the raw waveform feature set using SincNet [23] and the handcrafted feature set using openSMILE emobase2010 [96], which are of size $[1 \times 2048]$ and $[1 \times 1582]$ respectively. These two sets of features are then fed into two separately trained deep-learning CNN classifiers with self attention. Each pipeline is a CNN consisting of three sets of Convolution 1D, Batch Normalization, and Pooling layers. The first convolution layer consists of 32 filters of size = 3 and stride = 1. The second convolution layer has 64 filters. Each of those filters is of size = 3 and stride = 1. The third convolution layer has 128 filters with filter size = 3 and stride = 1. All the pooling layers other than the final pooling layer have filter size = 2 with stride = 2 and perform max pooling. The final pooling layer performs global average pooling. Besides, all the convolution layers use ReLU activation functions. We show the convolution1D, batch normalization, and max pooling layers as a single block termed L in Figure 4.3.

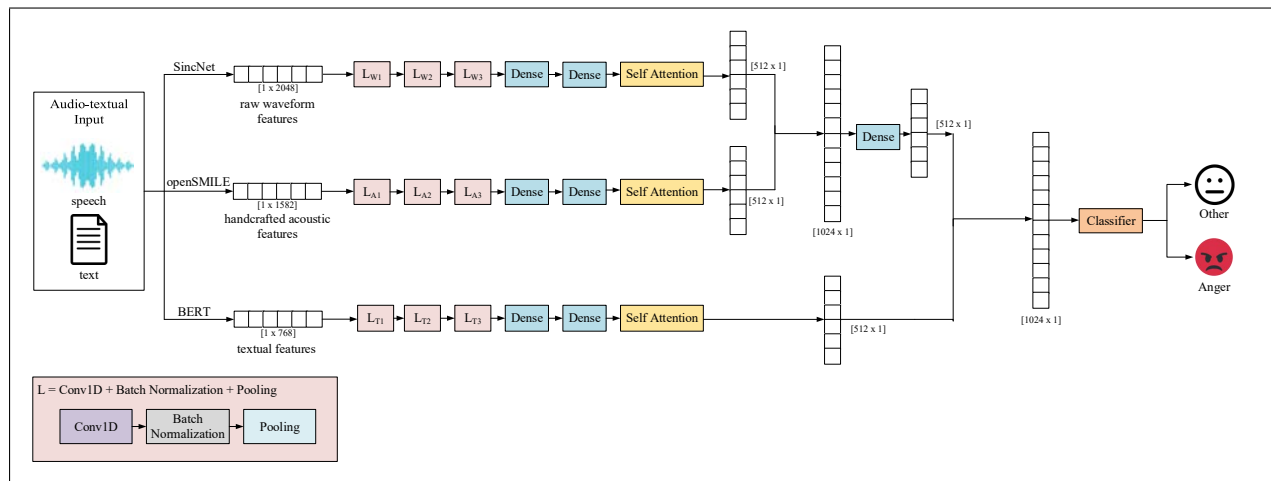


Figure 4.3: Traditional or Offline Anger Detector, **OAD**

Additionally, each pipeline contains two dense layers with 512 nodes each. The dense layers use the ReLU activation function. The output of the last dense layer is passed through a self attention layer. Then, the two output vectors of the CNN pipelines of size $[512 \times 1]$ are concatenated to form a vector of size $[1024 \times 1]$. This acts as the input of a dense layer with 512 neurons which acts as the final audio classifier. The output of the final audio classifier is of size $[512 \times 1]$.

Furthermore, we extract textual features of size $[1 \times 768]$ using BERT base [97]. Then, it is passed through a CNN pipeline with three blocks of convolution, batch normalization, and pooling layers. The first, second, and third convolution layers have 32, 64, and 128 filters respectively,

and use ReLU activation. Each of those filters is of size = 3 and stride = 1. All the pooling layers other than the final one have filter size = 2 with stride = 2 and perform max pooling. The final pooling layer performs global average pooling. Finally, the pipeline contains two dense layers with ReLU activation where each has 512 nodes. The output of the last dense layer is fed into a self attention layer which gives the output of the text classifier. The final text classifier output is of size $[512 \times 1]$.

The outputs from the audio and text classifiers are then fused in a mid-level fusion scheme and fed into a classifier. As the fusion classifier, we use an ensemble mechanism. In the experimentation section, we show the post-fusion performance comparison between different classifiers. We use XGBoost as the final classifier since it outperforms the baseline models including several ensemble mechanisms. Using this classifier, we classify the input into two classes: Anger and Other.

4.2 Experimental Evaluation

In this section, we describe the datasets used for the evaluation of our model. Besides, we discuss different baseline feature sets and fusion combinations along with the particular feature set and fusion combination used by our model. Moreover, we describe various baseline classifiers and baseline post-fusion classifiers.

4.2.1 Dataset Description

We use two datasets: the Bengali call-center dataset and the English IEMOCAP dataset. Both datasets contain conversational audio between two speakers. The datasets are discussed as follows.

The Bengali Call-Center Dataset

The Bengali call-center dataset contains dyadic conversations between the caller and the callee. The conversations are completely spontaneous conversations between real callers and agents. There are a total of 337 calls with a total duration of approximately 33 hours. Each call consists of an average of around 71 utterances and the average conversation time is around 5 minutes 53 seconds. There are a total of 23,766 utterances in the data set, with an average duration of approximately 4.99 seconds. The percentage of anger in the Bangla call-center dataset is shown in Table 4.1.

Since this dataset is pre-recorded and completely dependent on real-life situations for its participants, the data is not gender-balanced. There are slightly more female agents than

Conversation Type	Anger	Other
Spontaneous Conversations	28%	72%

Table 4.1: Percentage of angry utterances in the Bengali Call-center dataset

male agents and more male callers than female callers. The percentages of male and female speakers in the Bangla call-center dataset are shown in Table 4.2.

Speaker Type	Male	Female
Agent	41%	59%
Caller	62%	38%

Table 4.2: Percentage of male and female speakers in the Bengali Call-center dataset

In this dataset, the conversations are not separated into utterances. The data is also unlabeled with no transcriptions. Therefore, we perform additional pre-processing steps to prepare the data for our model. We transcribe the audio files using the paid version of Bengali Google Speech-to-Text [26]. Since the automatic transcriptions are not completely accurate, we perform manual corrections by checking with the audio clip. Besides, Google Speech-to-Text provides automated sentence segmentation by providing the start and end times of each utterance. We slightly correct those start and end durations and use them to separate the conversations into utterances.

Next, the utterances are labeled into two classes: anger and other, by listening to the audio clip. Three annotators labeled the data: one woman and two men. The annotators discussed the criteria for labeling in a common session and discussed some of the examples. Each utterance was labeled by at least two annotators. In cases where the labels assigned by the annotators were separate, a third annotator was consulted and majority voting was used to label the utterance. In cases of unclear utterances or noisy utterances, no labels could be assigned. Therefore, we discard the unlabeled utterances.

The IEMOCAP Dataset

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [27] is widely used in the field of emotion recognition. Since it contains conversational audio, visual information, and textual transcriptions, it can be used for multimodal emotion recognition. In our work, we make use of audio clips and transcriptions. In this dataset, the conversational audio is already broken down into utterances, and utterance-wise transcriptions are provided. There are a total of 10,039 utterances. Each utterance has an average length of 4.77 ± 3.34 seconds and was labeled by at least three annotators. Since the data is already pre-labeled, we do not need to perform additional operations in order to prepare the data for our classifier.

This database consists of conversational speech recorded by ten actors in dyadic sessions. There are fully scripted as well as spontaneous conversations. However, the spontaneous conversations

are generated by improvised hypothetical scenarios. While this situation is close to spontaneous conversations, it is still a bit different from the Bengali call-center dataset due to introducing hypothetical scenarios instead of real life conditions. Besides, due to being recorded in studio settings, the audio files are less noisy compared the Bengali dataset. IEMOCAP consists of approximately twelve hours of dyadic conversations. The percentages of anger in spontaneous and scripted conversations are shown in Table 4.3.

Conversation Type	Anger	Other
Scripted Conversations	7%	93%
Spontaneous Conversations	23%	77%
All Conversations	13%	87%

Table 4.3: Percentage of angry utterances in the IEMOCAP dataset

Additionally, the conversations are organized so that each conversation has a female and a male speaker. This ensures that the data is balanced in terms of gender. Besides, this dataset consists of conversational speech instead of isolated speech and is more likely to imitate real-life scenarios due to the presence of contextual information. The percentages of male and female speakers in spontaneous and scripted conversations are shown in Table 4.4.

Conversation Type	Male	Female
Scripted Conversations	50%	50%
Spontaneous Conversations	50%	50%

Table 4.4: Percentage of male and female speakers in the IEMOCAP dataset

4.2.2 Baseline Methods: Feature Sets and Fusion Combinations

In this section, we describe different feature sets and fusion combinations which are compared to our model. The features include audio (both handcrafted features and raw waveform features) and textual features. For all feature sets and fusion combinations, we use openSMILE emobase2010 [96] to extract handcrafted audio features, SincNet [23] to extract raw waveform features, and BERT base [97] along with BanglaBERT [98] to extract textual features.

Audio Features

This combination only uses audio features. Since textual features are not included, there is no concept of fusion in this case. We consider both handcrafted acoustic features and raw waveform features. Additionally, this scheme does not include gender information.

Audio Features (with gender)

Only audio features are used in this combination. This scheme does not require any fusion as textual features are not included here. It also incorporates gender information by using a gender classifier to classify the speaker of the utterance into male or female. There are two separately trained OADs in this scheme. According to the output of the gender classifier, we use the utterance as input to the appropriate OAD.

Audio and Text Features - Early Fusion

This combination uses both audio and textual features. However, instead of using separate pipelines for audio and text, this scheme merges the feature sets in early fusion to form a single feature set. Therefore, we do not require any fusion classifier here. Besides, this scheme does not include gender information.

Audio and Text Features - Early Fusion (with gender)

This combination makes use of both audio and textual attributes. Instead of using separate pipelines for audio and text, this scheme merges the feature sets in early fusion to form a single feature set. As a result, we do not need a fusion classifier in this case. It is similar to our model concerning incorporating gender information and uses a gender classifier to classify the gender of the speaker of the utterance. Then, the utterance is fed as input into the OAD corresponding to the gender of the speaker.

Audio and Text Features - Mid-level Fusion

This combination uses both audio and textual features and employs distinct classifiers for audio and textual features, which are then fused using a mid-level fusion scheme. A fusion classifier is later used to predict anger. This combination is very similar to our model discussed in Section 4.1.5. The only difference between this one and our model is the absence of gender information.

Audio (with gender) and Text Features - Mid-level Fusion

This is the feature set and fusion combination for our model. Our model uses separate classifiers for audio and textual features which are then fused following a mid-level fusion scheme. Later, a fusion classifier is used to predict anger. It also incorporates gender information by using a gender classifier to classify the speaker of the utterance into male or female. There are two OADs in this scheme and the utterance is used as input to the appropriate OAD according to the output of the gender classifier.

4.2.3 Baseline Methods: Classifiers

In this section, we describe different baseline classification schemes which are compared to our model. The baseline classifiers used are decision trees, random forests, k-nearest neighbours (k-NNs), support vector machines (SVMs), deep neural networks (DNNs), long short-term memory (LSTM) networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs). The classifiers are discussed in detail in Section 2.2. These baselines are derived from different existing works. However, we need to adapt these baselines since our feature sets and datasets are different from the existing works. Additionally, for the IEMOCAP dataset, we directly compare our model to state-of-the-art emotion recognition models for the IEMOCAP dataset by Xu et al. [102], Scotti et al. [103], and Priyasad et al. [19].

Decision Tree

Shami et al. [104] use decision trees to detect emotion from speech. As a baseline for our offline anger classification task, we use a decision tree with gini impurity and best split for each pipeline and also as the final classifier after fusion.

Random Forest

Zaman et al. [68] use random forests to detect emotion from speech. For each pipeline, we use a random forest classifier with 100 decision trees. In cases where a fusion classifier is required, we use another random forest classifier with 100 decision trees to obtain the output anger labels.

k-Nearest Neighbour (k-NN)

K-Nearest Neighbours have been used by many existing studies such as Zaman et al. [68], Shami et al. [104], etc. to detect emotion from speech. We use k-NN classifiers with $k = 5$ neighbours with uniform weights for each pipeline and also as the final classifier. All the k-NN models use the KDTree algorithm to compare the nearest neighbours.

Support Vector Machine (SVM)

Related works by Khalil et al. [6], Nomoto et al. [16], and Polzehl et al. [12] detect anger from audio-only or audio-textual data using SVMs. As our baseline, for each pipeline, we use an SVM classifier. In cases where a fusion classifier is required, we use another SVM model to obtain the output anger labels. All SVM models use polynomial degree= 2 kernel function.

Deep Neural Network (DNN)

Related work by Abu Shaqra et al. [1] and Shahriar et al. [20] use DNNs for emotion recognition. Our pipeline configuration is identical to the classifier used by Shahriar et al. [20]. We use fully connected layers or dense layers with the ReLU activation function as the primary classifier here. A detailed description of dense layers can be found in Section 2.2.2. In our baseline DNN, each pipeline consists of three dense layers with 256 neurons each with ReLU activation. After fusion, we use a dense layer with 512 neurons and ReLU activation and a subsequent softmax layer to obtain the final output.

Recurrent Neural Network (RNN)

Majumdar et al. [105] and Hasan et al. [14] use RNNs to detect emotion from speech. In each pipeline, we use two simple RNN layers with 256 neurons each and two subsequent dense layers with 256 neurons each. In cases where a fusion classifier is required, we use a dense layer with 512 neurons with ReLU activation and a subsequent softmax layer to obtain the final anger classification.

Long Short-Term Memory Network (LSTM)

Research studies by Deng et al. [3], Parry et al. [67], and Cai et al. [74] use LSTMs for emotion or anger detection. In our task, for each pipeline, we use two LSTM layers with 256 neurons each and two subsequent dense layers with ReLU activation and 256 neurons each. In cases where a fusion classifier is required, we use a dense layer of 512 neurons and a softmax layer to obtain the output anger labels.

Convolutional Neural Network (CNN)

Many existing works, such as the studies by Sultana et al. [24], Deng et al. [3], Parry et al. [67], and Cai et al. [74], use CNNs for emotion recognition or anger recognition. For each of the three pipelines, we use three blocks consisting of one conv1D layer, one batch normalization layer, and one pooling layer each. The final pooling layer performs global average pooling and the others perform max pooling. Afterwards, we use two subsequent dense layers in each pipeline. The configurations of the layers are the same as the configurations of the layers in our final model discussed in Subsection 4.1.5. The major difference between this model and our model is the absence of self attention and ensemble fusion classifier. In cases where a fusion classifier is required, we use a dense layer with 512 neurons with ReLU activation and a softmax layer to obtain the output anger labels.

State-of-the-Art Models

While comparing with different classifiers, we adapted those to be more suited to our particular feature sets and datasets. Additionally, for the IEMOCAP dataset, we use a separate set of experiments to compare our approach with the exact configurations of different state-of-the-art works on emotion recognition by Xu et al. (2021) [102], Scotti et al. (2021) [103], and Priyasad et al. (2020) [19] who validate their performances on the IEMOCAP dataset. Since these works [19, 102, 103] evaluate their performances on the IEMOCAP dataset, we can directly compare our performance with theirs without any modifications to their work. These works are similar to ours since they also use CNN architectures. Xu et al. (2021) [102] use a multi-head attention mechanism with a convolutional neural network (ACNN), termed Head Fusion. Their proposed modality is audio-only. Scotti et al. (2021) [103] perform audio-textual emotion recognition using PATHOSnet, a convolution neural network-based model. Additionally, Priyasad et al. (2020) [19] recognize emotion with the help of a CNN-based attention-driven fusion mechanism in the audio-textual domain.

4.2.4 Baseline Methods: Fusion Classifiers

In our model, we vary the fusion classifier to observe which classifier performs the best. We compare XGBoost to several baseline models for online anger classification, including Decision Tree, Random Forest, Adaboost, and Gradient Boost. The classification algorithms are discussed in detail in Sections 2.2.1 and 2.2.3. In particular, we use a decision tree with gini impurity and best split as the fusion classifier. Besides, we use a random forest classifier with 100 decision trees as the fusion classifier, where the decision trees use gini impurity and best split.

4.3 Results and Discussions

In this section, we evaluate the performance of OAD in comparison to baseline models including traditional machine learning models and deep learning models for the Bengali call-center dataset and the IEMOCAP dataset. We also evaluate the performances of our proposed feature set and fusion combination along with baseline feature sets and fusion combinations for all models. Additionally, we compare the performances of different fusion classifiers coupled with OAD. Since there is a class imbalance in the datasets, we adopt the F1 score as the performance metric.

4.3.1 Results: Bengali Call-Center Dataset

We perform offline anger classification using the feature sets and models discussed in Subsections 4.2.2 and 4.2.3. Table 4.5 shows the F1 scores for offline anger for different combinations of

classification models and feature sets for the Bengali call-center dataset. We choose XGBoost as the fusion classifier for OAD. For all feature sets and fusion combinations, OAD achieves the best F1 scores. CNN is a close second, with slightly worse F1 scores in all cases. We find that Decision Tree and k-NN especially struggle in cases of audio-only feature sets, with F1 scores below 65%.

Features	DT	RF	k-NN	SVM	DNN	RNN	LSTM	CNN	OAD
Audio	0.602	0.667	0.634	0.683	0.684	0.699	0.702	0.726	0.763
Audio (with gender)	0.625	0.682	0.653	0.702	0.713	0.714	0.717	0.746	0.792
Audio + Text: Early Fusion	0.627	0.683	0.658	0.709	0.714	0.715	0.723	0.752	0.795
Audio + Text: Early Fusion (with gender)	0.638	0.688	0.669	0.712	0.716	0.727	0.728	0.754	0.808
Audio + Text: Mid-Level Fusion	0.679	0.720	0.685	0.739	0.751	0.755	0.765	0.794	0.823
Audio (with gender) + Text: Mid-Level Fusion	0.713	0.734	0.732	0.767	0.778	0.784	0.797	0.810	0.855

Table 4.5: F1 score of offline anger for different feature sets and models for the Bengali call-center dataset

The results also validate our claims regarding textual features, mid-level fusion, and gender information. We find that mid-level fusion gives better F1 scores than early fusion in all cases. Moreover, including gender information outperforms the comparable feature sets that do not include gender information since men and women vary in using their voices to express anger. Additionally, textual features play a significant role in anger detection. This is why feature sets that include textual features perform significantly better than the ones that do not. Consequently, for all classification schemes, our chosen feature set and fusion combination: Audio (with gender) + Text: Mid-level Fusion performs the best. The feature sets and fusion combinations Audio + Text: Mid-level Fusion and Audio + Text: Early Fusion (with gender) also perform well, with F1 scores over 80% in the case of OAD. We achieve the best F1 score for the combination of OAD and Audio (with gender) + Text: Mid-level Fusion, which is 85.5%, outperforming the best performing baseline, CNN by 4.5%.

Features	Decision Tree	Random Forest	AdaBoost	Gradient Boost	XGBoost
Audio (with gender) + Text: Mid-Level Fusion	0.835	0.846	0.849	0.852	0.855

Table 4.6: F1 score of offline anger for different classifiers after fusion for the Bengali call-center dataset

Additionally, we compare the performances of different baseline classifiers with XGBoost. The

baseline classifiers are discussed in Subsection 4.2.4. In all cases, we use OAD as the classifier and Audio (with gender) + Text: Mid-level Fusion as the feature set and fusion combination. The F1 scores of offline anger for different fusion classifiers for the Bengali call-center dataset can be found in Table 4.6. We see that XGBoost performs better than the other fusion classifiers, achieving an F1 score of 85.5%, which is 0.3% better than Gradient Boost, the best performing baseline. Therefore, we choose XGBoost as the fusion classifier for OAD.

4.3.2 Results: IEMOCAP Dataset

For the IEMOCAP dataset, we perform similar experiments by varying the feature sets and classification schemes. For different combinations of classification models and feature sets, we observe the F1 scores for offline anger for the IEMOCAP dataset in Table 4.7. Similar to the experiments on the Bengali call-center dataset, we choose XGBoost as the fusion classifier for OAD. We see that OAD achieves the best F1 scores for all feature sets and fusion combinations. CNN and LSTM perform moderately well, with F1 scores over 85% for our chosen feature set and fusion combination. Decision tree, k-NN, and Random forest perform worse, with F1 scores below 80% for all feature sets and fusion combinations.

Features	DT	RF	k-NN	SVM	DNN	RNN	LSTM	CNN	OAD
Audio	0.631	0.726	0.693	0.742	0.743	0.758	0.761	0.785	0.822
Audio (with gender)	0.654	0.741	0.712	0.761	0.773	0.772	0.775	0.805	0.851
Audio + Text: Early Fusion	0.656	0.742	0.717	0.768	0.773	0.774	0.782	0.811	0.854
Audio + Text: Early Fusion (with gender)	0.667	0.747	0.728	0.771	0.775	0.786	0.786	0.813	0.867
Audio + Text: Mid-Level Fusion	0.728	0.779	0.744	0.798	0.809	0.814	0.824	0.853	0.882
Audio (with gender) + Text: Mid-Level Fusion	0.752	0.793	0.791	0.826	0.837	0.843	0.856	0.869	0.914

Table 4.7: F1 score of offline anger for different feature sets and models for the IEMOCAP dataset

The performances of schemes that incorporate the mid-level fusion technique are better than the performance of methods with early fusion for all anger detection models. Besides, for all classification models, the methods which include gender information perform better than the methods which do not. Moreover, our experiments show the importance of incorporating textual features in detecting anger since all feature sets without textual information perform worse than the feature sets with textual features. Consequently, our chosen feature set and fusion combination: Audio (with gender) + Text: Mid-level Fusion performs the best with all

classification models, with a 91.4% F1 score, outperforming the best performing baseline CNN by 4.5%.

Model	Modality	F1 score of anger
Head Fusion (Xu et al. (2021) [102])	A	75.8%
PATHOSnet (Scotti et al. (2021) [103])	A+T	84.9%
Attention-driven fusion F-I (Priyasad et al. (2020) [19])	A+T	86.1%
Attention-driven fusion F-II (Priyasad et al. (2020) [19])	A+T	87.0%
Attention-driven fusion F-III (Priyasad et al. (2020) [19])	A+T	86.8%
Our Approach: Audio-only	A	82.2%
Our Approach: OAD	A+T	91.4%

Table 4.8: F1 score of offline anger for different state-of-the-art models for the IEMOCAP dataset

Moreover, we use a separate set of experiments shown in Table 4.8 to directly compare our proposed method with the exact configurations of different state-of-the-art works on emotion recognition. These state-of-the-art models validate their performance on the IEMOCAP dataset. As a performance metric, we use the F1 score of anger. We show that our audio-only and audio-textual approaches outperform the corresponding baselines. Priyasad et al. [19] perform the second-best, achieving an F1 score of 87.0%.

Features	Decision Tree	Random Forest	AdaBoost	Gradient Boost	XGBoost
Audio (with gender) + Text: Mid-Level Fusion	0.885	0.905	0.908	0.912	0.914

Table 4.9: F1 score of offline anger for different classifiers after fusion for the IEMOCAP dataset

Additionally, we compare the performance of XGBoost with different baseline fusion classifiers. OAD is used as the anger classifier in all cases. Moreover, we use Audio (with gender) + Text: Mid-level Fusion as the feature set and fusion combination. We see that XGBoost performs better than the other fusion classifiers, achieving an F1 score of 91.4%, outperforming Gradient Boost by 0.2%. Gradient Boost also performs comparatively well, with more than 91% F1 score. Decision tree has the poorest performance, seconded by random forest. The F1 scores of offline anger for different fusion classifiers for the IEMOCAP dataset can be observed in Table 4.9.

4.3.3 Discussions and Limitations

For both the spontaneous Bengali call-center dataset and the acted IEMOCAP dataset, we compare our proposed Offline Anger Detector (OAD) with alternative baseline classification algorithms. We show that, in both datasets, our model consistently outperforms the baseline methods, regardless of language or domain. Additionally, we compare our proposed combination of feature set and fusion technique with different combinations of feature sets and fusion

techniques for our model and the baseline models. We demonstrate that our chosen combination of feature set and fusion technique performs better than the baselines. Besides, we show that incorporating gender information improves the performance of offline anger detection for all models, feature sets, and fusion techniques. Furthermore, we compare several fusion classifiers and find the best-performing fusion classifier.

Our work on offline anger detection has a few limitations. Training data are scarce in terms of quantity and variety. In addition, since we are working with real-life, spontaneous telephone conversations, the audio quality is not as good as that obtained from a studio setting. There are some missed utterances in the dataset due to noise or both of the speakers speaking together. These reduce the size of the dataset even further. A large amount of training data might have helped the model perform better. In real-life situations, the data can be skewed, e.g., there might be few instances of hot anger in a particular data set. Furthermore, spontaneous conversations contain very few instances of strong emotion in most cases.

Besides, we require a gender classifier to incorporate gender information into offline anger classification. The gender classifier might have performed better with a larger dataset. Moreover, we perform speech-to-text transcriptions with the help of Google speech-to-text. The transcriptions are quite faulty and need manual corrections, which is highly time-consuming. However, we perform manual edits to achieve higher performance for offline anger detection. Besides, spontaneous conversational Bengali contains regional words, incomplete sentences, wrong pronunciation, and wrong grammatical structure. These errors negatively impact the performance of our model. However, to keep the transcriptions as authentic as possible, we avoid correcting these errors.

Chapter 5

Online Anger Detection

In this chapter, we propose a novel anger detection mechanism from previous utterances of audio-textual conversations, which is termed online anger detection. At first, we explore our model for online anger detection. Then, we describe our datasets and baseline methods. Afterwards, we discuss our findings and note the limitations.

5.1 Methodologies

Firstly, we describe the intuitions behind our approach. Secondly, we explain how to model the input and output data in terms of utterances. Finally, we discuss our early or online anger detection model in detail, which is termed **Early/Online Anger Detector (EAD)**.

5.1.1 Intuitions behind the Proposed Approach

Early or online anger detection is very different from offline anger detection since it is mainly a multivariate forecasting problem instead of a simple classification. Besides, in order to incorporate historical information during the early detection of anger, we need to consider multiple previous utterances at a time. Additionally, there can be missing utterances due to noise and the output data can come from either of the two participants. Therefore, online audio-textual anger detection is a complex problem that requires special attention.

To the best of our knowledge, there are no existing works on online audio-textual anger recognition. Related works by Mongkolnavin et al. [21] and Shahriar et al. [20] deal with audio-only or audio-visual data, which do not work well for cold anger detection. Moreover, they only consider handcrafted features, whereas raw waveform features perform well for these tasks. Besides, they consider the utterances of a single speaker in a dyadic conversation despite the influence of the utterances of both speakers on the emotional flow of participants [22].

Therefore, we propose an ensemble-based deep learning approach for online audio-textual anger recognition which incorporates textual features, the utterances of both participants, and sinc-based raw speech waveform features alongside handcrafted acoustic features.

5.1.2 Modeling the Input and Output Data

In our work, we predict utterance-level online anger for dyadic conversations. To determine the start and end time of an utterance, we use Voice Activity Detection (VAD). As a result, the utterances can be of varying lengths. Furthermore, we remove the noisy segments where both speakers are speaking at the same time by detecting simultaneous activity in both channels.

There are three utterance parameters: number of input utterances i , number of skipped utterances s , and number of output utterances o . Our objective is to use the i input utterances to obtain the anger labels of the o output utterances. Additionally, during online anger prediction in the real world, there can be missing utterances due to noise or distortion. Therefore, to simulate real world data, we consider s skipped utterances.

In this work, we consider the influence of one speaker and both speakers in a dyadic conversation. The schemes are described in the following sections.

Modeling the Data for a Single Participant

We show the modeling of data for a single participant in a dyadic conversation between two speakers: **A** and **B**. In the case of a single participant, we predict online anger by considering the utterances of participant **A** only. All the utterances of participant **B** are discarded in this case. Let the sequence of all of the n utterances of participant **A** be \mathbf{U}_A , which can be expressed as follows.

$$\mathbf{U}_A = (U_{A,1}, U_{A,2}, \dots, U_{A,n})$$

The sequence of all of the n utterance labels \mathbf{L}_A can be expressed as follows.

$$\mathbf{L}_A = (L_{A,1}, L_{A,2}, \dots, L_{A,n})$$

If the input utterances begin at the first utterance $U_{A,1}$, then the sequence of utterances of participant **A**, \mathbf{U}_A can be expressed as follows.

$$\mathbf{U}_A = (U_{A,1}, U_{A,2}, \dots, U_{A,i}, U_{A,i+1}, \dots, U_{A,i+s}, U_{A,i+s+1}, \dots, U_{A,i+s+o}, U_{A,i+s+o+1}, \dots, U_{A,n})$$

If the sequences of input, skipped, and output utterances are \mathbf{I}_A^U , \mathbf{S}_A^U , and \mathbf{O}_A^U respectively, and

the sequence of output labels is \mathbf{O}_A^L then the sequences can be represented as follows.

$$\begin{aligned}\mathbf{I}_A^U &= (U_{A,1}, U_{A,2}, \dots, U_{A,i}) \\ \mathbf{S}_A^U &= (U_{A,i+1}, U_{A,i+2}, \dots, U_{A,i+s}) \\ \mathbf{O}_A^U &= (U_{A,i+s+1}, U_{A,i+s+2}, \dots, U_{A,i+s+o}) \\ \mathbf{O}_A^L &= (L_{A,i+s+1}, L_{A,i+s+2}, \dots, L_{A,i+s+o})\end{aligned}$$

Therefore, the mapping of input feature data to labels will be, $\mathbf{I}_A^U \rightarrow \mathbf{O}_A^L$. The data modeling scheme for a single participant is shown in Figure 5.1.

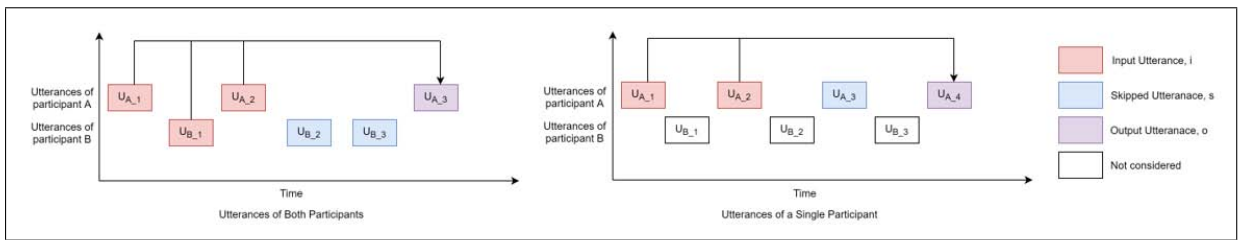


Figure 5.1: Input utterances, output utterances, and skipped utterances explained

In Figure 5.1, the input utterances start from the first utterance of participant A, $U_{A,1}$. All the utterances of participant B are discarded. Here, the utterances and labels for participant A are shown as follows.

$$\begin{aligned}\mathbf{U}_A &= (U_{A,1}, U_{A,2}, U_{A,3}, U_{A,4}) \\ \mathbf{L}_A &= (L_{A,1}, L_{A,2}, L_{A,3}, L_{A,4})\end{aligned}$$

In this example, the number of input utterances $i = 2$, the number of skipped utterances $s = 1$, and the number of output utterances $o = 1$. Therefore, the sequences \mathbf{I}_A^U , \mathbf{S}_A^U , \mathbf{O}_A^U , and \mathbf{O}_A^L can be expressed as, $\mathbf{I}_A^U = (U_{A,1}, U_{A,2})$, $\mathbf{S}_A^U = (U_{A,3})$, $\mathbf{O}_A^U = (U_{A,4})$, and $\mathbf{O}_A^L = (L_{A,4})$ respectively. The mapping of input utterances to labels is $(U_{A,1}, U_{A,2}) \rightarrow (L_{A,4})$.

Modeling the Data for Both Participants

When both participants are considered, we consider the sequence of all utterances in the conversation, \mathbf{U} , where the number of utterances is n . Then, the sequence of all utterances \mathbf{U} and the sequence of all labels \mathbf{L} can be represented as follows.

$$\begin{aligned}\mathbf{U} &= (U_1, U_2, \dots, U_n) \\ \mathbf{L} &= (L_1, L_2, \dots, L_n)\end{aligned}$$

If the sequence of input, skipped, and output utterances are \mathbf{I}^U , \mathbf{S}^U , and \mathbf{O}^U respectively, and the sequence of output labels is \mathbf{O}^L then the sequences can be represented as follows.

$$\begin{aligned}\mathbf{I}^U &= (U_1, U_2, \dots, U_i) \\ \mathbf{S}^U &= (U_{i+1}, U_{i+2}, \dots, U_{i+s}) \\ \mathbf{O}^U &= (U_{i+s+1}, U_{i+s+2}, \dots, U_{i+s+o}) \\ \mathbf{O}^L &= (L_{i+s+1}, L_{i+s+2}, \dots, L_{i+s+o})\end{aligned}$$

Therefore, the mapping of input feature data to labels will be, $\mathbf{I}^U \rightarrow \mathbf{O}^L$. In the example of Figure 5.1, the number of input utterances $i = 3$, the number of skipped utterances $s = 2$, and the number of output utterances $o = 1$. Therefore, the sequences \mathbf{I}^U , \mathbf{S}^U , \mathbf{O}^U , and \mathbf{O}^L can be expressed as, $\mathbf{I}^U = (U_{A,1}, U_{B,1}, U_{A,2})$, $\mathbf{S}^U = (U_{B,2}, U_{B,3})$, $\mathbf{O}^U = (U_{A,4})$, and $\mathbf{O}^L = (L_{A,4})$ respectively. The mapping of input utterances to labels is $(U_{A,1}, U_{B,1}, U_{A,2}) \rightarrow (L_{A,4})$.

5.1.3 Feature Extraction

For online anger detection, we use both audio and text features. For audio, we use handcrafted acoustic features obtained from openSMILE emobase2010 [96] and raw waveform features from SincNet [23]. For textual features, we use BERT base [97] for the IEMOCAP dataset and BanglaBERT [98] for the Bengali call-center dataset. The feature extraction schemes are discussed in detail in Subsections 4.1.3 and 4.1.4.

5.1.4 Online Anger Detection

For online anger detection, the conversational audio is split into utterances, which are further split into $250ms$ segments. For each segment, we obtain the response of the classifier, and the response for the whole utterance is found by adding the responses obtained from each segment to get the final classification scores, similar to SincNet [23].

We use two types of features in this work: audio and text. The audio features can be further split into two types: raw waveform features and handcrafted acoustic features. For each of the three types of feature sets, we use a separate spatio-temporal transformer model Spacetimeformer by Grigsby et al. [106]. The outputs from the two audio pipelines are merged which gives the output of the final audio classifier. Then, the audio classifier output is fused with the output from the text classifier in a mid-level fusion scheme. The Spacetimeformer model used in EAD is discussed as follows.

Spacetimeformer

Grigsby et al. [106] propose Spacetimeformer, which is a model for spatio-temporal multivariate forecasting using long-range transformers. Spacetimeformer provides a more efficient solution in the case of long-input domains compared to the original Transformer [48] architecture.

Transformer’s input data follows the pattern (x_t, \mathbf{y}_t) , which are concatenated vectors of multiple variable values per timestep. The spatio-temporal embedding discussed in Spacetimeformer divides the inputs (x_t, \mathbf{y}_t) into a subsequence of tokens. If the number of variables being modeled is V , then the embedding is a sub-sequence of $\{(x_t, y_t^i), \dots, (x_t, y_t^V)\}$ tokens. This embedding represents each node of the spatio-temporal graph as a separate token and allows us to understand complex relationships between variables.

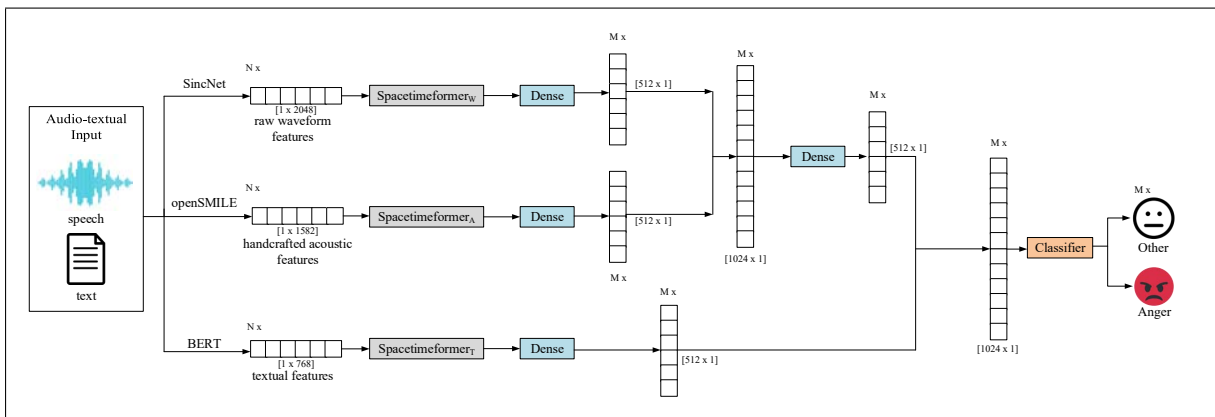
Besides, positional embeddings used in the original transformer only maintain the relative order of data, whereas long-term patterns in forecasting problems can depend on the actual time window in which an event occurs. Spacetimeformer uses a Time2Vec layer to solve this problem by representing periodic relationships that are required for accurate predictions. The spatio-temporal encoding is then concatenated with this time embedding, which is fed through a feed-forward network in order to get the input of the transformer model. This is the Value & Time Embedding. Besides, the input also includes a given embedding which indicates whether a particular value is a context value or needs prediction. Moreover, the input includes a variable embedding or lookup-table embedding which maps the index of each time series to a spatial representation. These pre-processed embeddings are then used as input to a standard transformer Encoder-Decoder network.

Spacetimeformer also incorporates several optimization schemes. They use the Performer FAVOR + attention scheme, which is a linear approximation of softmax attention and incorporates a random kernel method. Moreover, they use Local Attention alongside Global Attention, which allows incorporating local bias which is especially important in cases of large V . In this scheme, a token first pays attention to every token in its sequence and then pays attention to every token in the global sequence. Besides, they improve performance by scaling the data by applying convolution to the full sequence of each variable independently, which halves their length. They are then recombined into a longer sequence to preserve the initial emphasis on the token ordering.

Online Anger Detector (EAD): The Complete Model

The complete model of our Early or Online Anger Detector (EAD) is shown in Figure 5.2.

If the number of input utterances, $i = N$, the number of skipped utterances, $s = K$, and the number of output utterances, $o = M$, then for each of the N input utterances, the raw waveform feature set of size $[1 \times 2048]$ and the handcrafted feature set of size $[1 \times 1582]$ are obtained using SincNet [23] and openSMILE emobase2010 [96] respectively. These two sets of features for N input utterances are then fed into two separately trained pipelines. Each pipeline consists of a spacetimeformer model and a dense layer with 512 nodes and ReLU activation function. The output of the spacetimeformer is then passed through the dense layer. Next, for each of the M output utterances, the two output vectors for the two types of feature sets are merged to form a vector of size $[1024 \times 1]$. This acts as the input of a dense layer with 512 neurons which acts as

Figure 5.2: Early or Online Anger Detector, **EAD**

the final audio classifier. The output of the final audio classifier is M vectors of size $[512 \times 1]$. Besides, for each of the N input utterances, we extract textual features of size $[1 \times 768]$. Then, it is passed through a spacetimeformer pipeline with a dense layer. The dense layer has 512 neurons with ReLU activation, which gives the output of the text classifier. The final text classifier outputs are M vectors of size $[512 \times 1]$.

The audio and text classifier outputs are then fused in a mid-level fusion scheme and fed into a classifier. We employ an ensemble model as the fusion classifier. In the experimentation section, we compare the post-fusion performance of various classifiers. As the final classifier, we use XGBoost since it outperforms the baseline models including several ensemble mechanisms. We classify each of the M output utterances into one of the two categories using this classifier: anger and other.

5.2 Experimental Evaluation

In this section, we mention the datasets used for the evaluation of our model. Besides, we discuss different baseline feature sets and fusion combinations along with the particular feature set and fusion combination used by our model. Moreover, we describe various baseline classifiers and baseline fusion classifiers.

5.2.1 Dataset Description

In this work, we use the Bangla call-center dataset and the English IEMOCAP dataset. The detailed dataset descriptions can be found in Subsection 4.2.1.

5.2.2 Baseline Methods: Feature Sets and Fusion Combinations

In this section, we describe different feature sets and fusion combinations which are compared to our model. The features include audio (both handcrafted features and raw waveform features) and textual features. For all feature sets and fusion combinations, we use openSMILE emobase2010 [96] to extract handcrafted audio features, SincNet [23] to extract raw waveform features and BERT base [97] along with BanglaBERT [98] to extract textual features.

Audio-only, without raw waveform (single)

This combination only uses handcrafted audio features. Since textual features are not included, there is no concept of fusion in this case. We use only one classifier pipeline to predict online anger. Additionally, this combination only considers the utterances of a single participant.

Audio-only, without raw waveform (both)

This combination does not use raw waveform features and textual features. As a result, there is no need for a fusion classifier. Besides, this scheme uses only one classifier pipeline to predict online anger. Moreover, it considers the utterances of both speakers in a dyadic conversation.

Audio-only, with raw waveform (single)

This combination uses both raw waveform features and handcrafted audio features. However, since textual features are not included, there is no concept of fusion in this case. We use two classifiers for the two types of audio features to predict online anger. Additionally, this combination only considers the utterances of a single participant.

Audio-only, with raw waveform (both)

This combination uses raw waveform features and handcrafted audio features. Nevertheless, it does not use textual features and does not require a fusion classifier. Besides, this scheme uses two classifiers for raw waveform features and handcrafted audio features. Moreover, it considers the utterances of both speakers in a dyadic conversation.

Audio + Text (single)

This combination uses both audio and textual features and employs distinct classifiers to obtain two different outputs for audio and textual features, which are then fused using a mid-level fusion scheme. A fusion classifier is later used to predict anger. This combination is very similar to our model for detecting online anger. The only difference between this one and our model is that this

scheme considers the utterances of a single participant while our model considers the utterances of both participants.

Audio + Text (both)

This is the feature set and fusion combination for our model. Our model uses separate classifiers for audio and textual features. The outputs from the audio and text classifiers are then fused following a mid-level fusion scheme. Later, a fusion classifier is used to predict anger. Our model considers the utterances of both speakers in a dyadic conversation.

5.2.3 Baseline Methods: Classifiers

In this section, we describe different baseline classification schemes which are compared to our model. The baseline classifiers used are deep neural networks (DNNs), long short-term memory (LSTM) networks, and bidirectional long short-term memory (Bi-LSTM) networks. The baseline models are discussed in Section 2.2.2. Additionally, for the IEMOCAP dataset, we use a separate set of experiments to compare our audio-only (single speaker) approach with different state-of-the-art models which evaluate their performances on the IEMOCAP dataset.

Deep Neural Network (DNN)

Related works by Abu Shaqra et al. [1] and Shahriar et al. [20] use DNNs for emotion recognition. Our pipeline configuration is identical to the classifier used by Shahriar et al. [20]. For each pipeline, we use three dense layers with 256 neurons each with ReLU activation. After fusion, we use a dense layer with 512 neurons and ReLU activation and a subsequent softmax layer to obtain the final output.

Long Short-Term Memory Network (LSTM)

Related works by Mongkolnavin et al. [21] and Shahriar et al. [20] use LSTMs for emotion or anger recognition. In our task, for each pipeline, we use two LSTM layers and two subsequent dense layers with ReLU activation and 256 neurons each. In cases where a fusion classifier is required, we use a dense layer of 512 neurons and a softmax layer to obtain the output anger labels.

Bidirectional Long Short-Term Memory Network (Bi-LSTM)

Shahriar et al. [20] and Sultana et al. [24] use Bi-LSTMs to detect emotion. We employ two Bi-LSTM layers and two following dense layers with ReLU activation and 256 neurons for each

pipeline in our work. In circumstances where a fusion classifier is required, we predict the output anger labels using a dense layer of 512 neurons and a softmax layer.

State-of-the-Art Models

Furthermore, we use a separate set of experiments to directly compare our audio-only (single speaker) approach with the exact configurations of different state-of-the-art models which evaluate their performance on the IEMOCAP dataset, considering audio features only. During the comparison, we only consider audio features since visual features are not always available and are outside the domain of our work. For the baselines and our approach, we only consider the utterances of a single speaker since this is how the baselines were modeled. Shahriar et al. [20] use DNNs, LSTMs and Bi-LSTMs to detect emotion from audio-visual data using the IEMOCAP dataset. Mongkolnavin et al. [21] use LSTMs to detect forthcoming anger from audio-only data.

5.2.4 Baseline Methods: Fusion Classifiers

In our model, we vary the fusion classifier to observe which classifier performs the best. For online anger classification, we compare XGBoost with several baseline models such as Decision Tree, Random Forest, AdaBoost, and Gradient Boost. A detailed discussion on the baseline models can be found in Subsection 4.2.4.

5.3 Results and Discussions

In this section, we evaluate the performance of EAD in comparison to baseline models for the Bengali call-center dataset and the IEMOCAP dataset. We also compare the performance of our proposed feature set and fusion combination along with baseline feature sets for all models. Additionally, we compare the performance of different fusion classifiers coupled with EAD. Furthermore, we vary the numbers of input, output, and skipped utterances and observe their effects on the performance of the models. Since there is a class imbalance in the datasets, we adopt the F1 score as the performance metric.

5.3.1 Results: Bengali Call-Center Dataset

For the Bengali call-center dataset, we perform online anger detection using the feature sets and models discussed in Subsections 5.2.2 and 5.2.3. Table 5.1 shows the F1 scores for online anger for different combinations of classification models and feature sets. For these experiments, the default values of the number of input, output, and skipped utterances are $i = 3$, $o = 1$, and

$s = 0$. The reasons for choosing these particular values are explained later along with a detailed discussion of the impacts of these parameters on the performance of online anger detection.

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.549	0.566	0.586	0.612
Audio-only, without raw waveform (both)	0.553	0.575	0.591	0.619
Audio-only, with raw waveform (single)	0.560	0.579	0.597	0.626
Audio-only, with raw waveform (both)	0.569	0.585	0.608	0.644
Audio + Text (single)	0.585	0.591	0.615	0.649
Audio + Text (both)	0.603	0.619	0.627	0.669

Table 5.1: F1 score of online anger for different feature sets and models for the Bengali call-center dataset ($i = 3, o = 1, s = 0$)

As the fusion classifier for EAD, we choose XGBoost. EAD achieves the best F1 scores for all feature sets and fusion combinations. We find that DNN and LSTM struggle the most, achieving F1 scores below 60% for all cases other than our chosen feature set and fusion combination.

The results also validate our claims regarding textual features, raw waveform features, and the utterances of both participants. Raw waveform features significantly improve the performance of online anger detection. Besides, considering the utterances of both speakers perform better than considering the utterances of only one speaker. Furthermore, textual features play a significant role in anger detection. As a result, feature sets that include textual features perform significantly better than feature sets that do not. Therefore, our chosen feature set performs the best for all classification models. For EAD, our chosen feature set and fusion combination achieve an F1 score of 66.9%, which outperforms the best performing baseline Bi-LSTM by 4.2%.

Features	Decision Tree	Random Forest	Adaboost	Gradient Boost	XGBoost
Audio + Text (both)	0.648	0.654	0.659	0.661	0.669

Table 5.2: F1 score of online anger for different classifiers after fusion for the Bengali call-center dataset ($i = 3, o = 1, s = 0$)

Besides, we compare the performance of different baseline fusion classifiers coupled with EAD. The baseline fusion classifiers are discussed in Subsection 5.2.4. In all cases, we use EAD as the classifier with our chosen feature set and fusion combination. The number of input, output, and skipped utterances are set at their default values with $i = 3, o = 1$, and $s = 0$. The F1 scores of online anger for different fusion classifiers for the Bengali call-center dataset can be found in Table 5.2. We see that XGBoost performs better than the other fusion classifiers, achieving an F1 score of 66.9% and outperforming Gradient Boost by 0.8%. Therefore, we choose XGBoost as the fusion classifier for the rest of our experiments.

Moreover, we vary the number of skipped utterances, s , to find its impact on the performance of online anger. The number of input and output utterances are set at their default values with

Number of skipped utterances, s	DNN	LSTM	Bi-LSTM	EAD
0	0.603	0.619	0.627	0.669
1	0.562	0.586	0.602	0.631
2	0.534	0.548	0.555	0.579
3	0.501	0.510	0.511	0.516

Table 5.3: Impact of the number of skipped utterances, s , on the F1 score of online anger for ($i = 3, o = 1$) for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset

$i = 3$ and $o = 1$. For all cases, we consider our chosen feature set and fusion combination with XGBoost as the fusion classifier for EAD. We see that EAD performs better than the baseline classifiers for all values of s . The impact of s on the performance of online anger can be seen in Table 5.3.

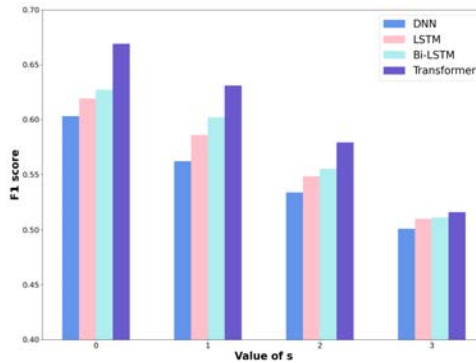


Figure 5.3: Impact of varying the number of skipped utterances, s , on the F1 score of online anger for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset

Naturally, keeping the number of skipped utterances, $s = 0$, gives the best performance since increasing the utterance steps between the input and the output utterances indicates that important information is lost. However, in the real world, there can be missing utterances due to noise or distortion. Therefore, we vary the value s to understand how many utterances can be dropped without affecting the performance by a significant margin. We find that increasing the value of s from 1 to 2 makes the F1 score drop below 60%. Therefore, we keep the number of skipped utterances, $s = 0$ for most of our experiments. The impact of varying the number of skipped utterances, s , on the performance of online anger detection is shown in Figure 5.3.

Additionally, we vary the numbers of input and output utterances to find the best combination for online anger detection. For all values of i and o , we consider our chosen feature set and fusion combination with EAD as the online anger classifier. We observe the impact of the values of input and output utterances in Table 5.4. We find that increasing the number of input utterances generally improves the F1 score. However, increasing the value of output utterances, o will naturally decrease performance. Unfortunately, due to resource constraints, we could not vary

Number of input utterances, i	Number of output utterances, o	
	1	2
1	0.628	0.619
2	0.647	0.638
3	0.669	0.656

Table 5.4: Impact of the number of input utterances, i and the number of output utterances, o , on the F1 score of online anger for EAD for audio and textual features considering the utterances of both speakers in dyadic conversations for the Bengali call-center dataset

the values of i and o too much. For most of our experiments, the default values of input and output utterances are $i = 3$ and $o = 1$.

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.510	0.542	0.546	0.586
Audio-only, without raw waveform (both)	0.514	0.546	0.551	0.599
Audio-only, with raw waveform (single)	0.517	0.549	0.558	0.605
Audio-only, with raw waveform (both)	0.534	0.551	0.562	0.619
Audio + Text (single)	0.543	0.554	0.564	0.624
Audio + Text (both)	0.551	0.576	0.585	0.628

Table 5.5: F1 score of online anger for the Bengali call-center dataset ($i = 1, o = 1, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.501	0.533	0.538	0.578
Audio-only, without raw waveform (both)	0.503	0.535	0.544	0.582
Audio-only, with raw waveform (single)	0.512	0.539	0.548	0.595
Audio-only, with raw waveform (both)	0.518	0.541	0.554	0.609
Audio + Text (single)	0.533	0.546	0.559	0.613
Audio + Text (both)	0.547	0.562	0.575	0.619

Table 5.6: F1 score of online anger for the Bengali call-center dataset ($i = 1, o = 2, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.529	0.561	0.573	0.595
Audio-only, without raw waveform (both)	0.532	0.564	0.578	0.612
Audio-only, with raw waveform (single)	0.538	0.566	0.585	0.614
Audio-only, with raw waveform (both)	0.552	0.569	0.592	0.631
Audio + Text (single)	0.569	0.573	0.595	0.636
Audio + Text (both)	0.584	0.602	0.609	0.647

Table 5.7: F1 score of online anger for the Bengali call-center dataset ($i = 2, o = 1, s = 0$)

Furthermore, for all combinations of the input and output utterances mentioned in Table 5.4, we show the performance of EAD and our chosen feature set and fusion combination alongside the performance of the baseline models and feature sets. The results of the experiments can be found

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.521	0.549	0.560	0.589
Audio-only, without raw waveform (both)	0.525	0.555	0.565	0.609
Audio-only, with raw waveform (single)	0.529	0.557	0.568	0.611
Audio-only, with raw waveform (both)	0.549	0.559	0.575	0.626
Audio + Text (single)	0.562	0.565	0.583	0.629
Audio + Text (both)	0.572	0.591	0.602	0.638

Table 5.8: F1 score of online anger for the Bengali call-center dataset ($i = 2, o = 2, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.538	0.565	0.585	0.610
Audio-only, without raw waveform (both)	0.541	0.569	0.589	0.613
Audio-only, with raw waveform (single)	0.553	0.572	0.595	0.617
Audio-only, with raw waveform (both)	0.560	0.575	0.602	0.635
Audio + Text (single)	0.582	0.586	0.609	0.639
Audio + Text (both)	0.591	0.613	0.625	0.656

Table 5.9: F1 score of online anger for the Bengali call-center dataset ($i = 3, o = 2, s = 0$)

in Tables 5.5, 5.6, 5.7, 5.8, and 5.9. For all values of input and output utterances, our chosen feature set gives the best performance. Similarly, for all cases, EAD performs better than the baseline models.

5.3.2 Results: IEMOCAP Dataset

Similarly, for the IEMOCAP dataset, we perform online anger detection using different feature sets and classification models. Table 5.10 shows the F1 scores for online anger for different combinations of classification models and feature sets. The values of the number of input, output, and skipped utterances are at their default values with $i = 3, o = 1, s = 0$ respectively. The reasons for choosing these particular values are explained in the previous section.

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.552	0.569	0.588	0.613
Audio-only, without raw waveform (both)	0.556	0.578	0.593	0.625
Audio-only, with raw waveform (single)	0.564	0.581	0.598	0.629
Audio-only, with raw waveform (both)	0.574	0.582	0.611	0.647
Audio + Text (single)	0.585	0.594	0.618	0.652
Audio + Text (both)	0.608	0.625	0.634	0.677

Table 5.10: F1 score of online anger for different feature sets and models for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)

For all combinations of feature sets, XGBoost is chosen as the fusion classifier for EAD. EAD achieves the best F1 scores for all feature sets and fusion combinations. We find that DNN struggles the most in terms of performance. For all models, our chosen feature set outperforms

all the baseline feature sets. For EAD, our chosen feature set and fusion combination achieve an F1 score of 67.7%, outperforming the best performing baseline Bi-LSTM by 4.3%. Similar to the results on the Bengali dataset, these results also demonstrate the importance of raw waveform features, textual features, and the utterances of both participants.

Model	Modality	F1 score of Anger
DNN (Shahriar et al. (2019) [20])	A	56.4%
LSTM (Shahriar et al. (2019) [20])	A	57.9%
Bi-LSTM (Shahriar et al. (2019) [20])	A	59.4%
LSTM (Mongkolnavin et al. (2020) [21])	A	58.1%
Our Approach: Audio-only (single)	A	62.9%

Table 5.11: F1 score of online anger for different state-of-the-art models for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)

Furthermore, we compare our audio-only approach with different state-of-the-art models in Table 5.11, considering audio features only. We use this separate set of experiments since the problem setting of the baselines is different from ours, and in most cases, the baselines use visual features too. Therefore, we adopt the exact configurations of their models with audio-only features for comparison. Mongkolnavin et al. [21] use LSTMs to detect forthcoming anger from audio-only data. In addition to LSTMs, Shahriar et al. [20] use DNNs and Bi-LSTMs to detect emotion from audio-visual data using the IEMOCAP dataset. During the comparison, we only consider audio features since visual features are outside the domain of our work. Our approach performs better than the baselines, outperforming the best baseline model Bi-LSTM by Shahriar et al. [20] by 3.5%.

Features	Decision Tree	Random Forest	Adaboost	Gradient Boost	XGBoost
Audio + Text: Both	0.651	0.659	0.663	0.668	0.677

Table 5.12: F1 score of online anger for different classifiers after fusion for the IEMOCAP dataset ($i = 3, o = 1, s = 0$)

Additionally, we compare the performance of different fusion classifiers. EAD is used as the classifier along with our chosen feature set and fusion combination. The numbers of input, output, and skipped utterances are also set at their default values. The F1 scores of online anger for different fusion classifiers for the Bengali call-center dataset can be found in Table 5.12. XGBoost performs the best, achieving an F1 score of 67.7% and outperforming Gradient Boost by 0.9%. Similar to the Bengali call-center dataset, XGBoost is chosen as the fusion classifier for the rest of our experiments.

Besides, we vary the number of skipped utterances, s for the IEMOCAP dataset. The number of input and output utterances are set at their default values. For all cases, we consider our chosen feature set and fusion combination with XGBoost as the fusion classifier for EAD. As

Number of skipped utterances, s	DNN	LSTM	Bi-LSTM	EAD
0	0.608	0.625	0.634	0.677
1	0.568	0.592	0.607	0.639
2	0.539	0.554	0.562	0.587
3	0.506	0.516	0.518	0.524

Table 5.13: Impact of the number of skipped utterances, s , on the F1 score of online anger for ($i = 3, o = 1$) for audio and textual features considering the utterances of both speakers in dyadic conversations for the IEMOCAP dataset

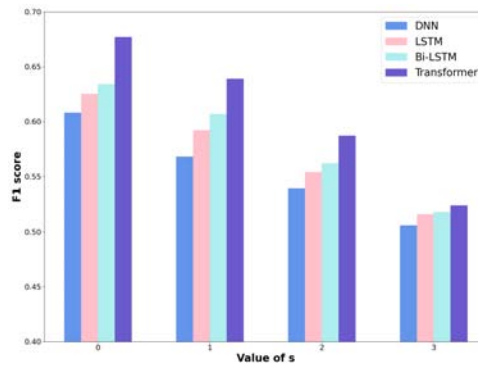


Figure 5.4: Impact of varying the number of skipped utterances, s , on the F1 score of online anger for audio and textual features considering both speakers for the IEMOCAP dataset

expected, EAD performs better than the baseline classifiers in all cases. The impact of s on the performance of online anger can be seen in Table 5.13. The impact of varying s on the performance of online anger detection is shown in Figure 5.4.

Number of input utterances, i	Number of output utterances, o	
	1	2
1	0.630	0.621
2	0.652	0.643
3	0.677	0.659

Table 5.14: Impact of the number of input utterances, i and the number of output utterances, o , on the F1 score of online anger for EAD for audio and textual features considering the utterances of both speakers in dyadic conversations for the IEMOCAP dataset

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.513	0.545	0.549	0.588
Audio-only, without raw waveform (both)	0.517	0.549	0.554	0.603
Audio-only, with raw waveform (single)	0.520	0.552	0.560	0.607
Audio-only, with raw waveform (both)	0.538	0.553	0.565	0.621
Audio + Text (single)	0.546	0.557	0.568	0.627
Audio + Text (both)	0.554	0.579	0.588	0.630

Table 5.15: F1 score of online anger for the IEMOCAP dataset ($i = 1, o = 1, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.504	0.536	0.542	0.580
Audio-only, without raw waveform (both)	0.506	0.539	0.547	0.585
Audio-only, with raw waveform (single)	0.515	0.541	0.551	0.598
Audio-only, with raw waveform (both)	0.521	0.543	0.556	0.611
Audio + Text (single)	0.537	0.549	0.562	0.616
Audio + Text (both)	0.549	0.565	0.578	0.621

Table 5.16: F1 score of online anger for the IEMOCAP dataset ($i = 1, o = 2, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.532	0.563	0.576	0.598
Audio-only, without raw waveform (both)	0.535	0.567	0.581	0.615
Audio-only, with raw waveform (single)	0.541	0.569	0.587	0.618
Audio-only, with raw waveform (both)	0.556	0.571	0.592	0.633
Audio + Text (single)	0.571	0.576	0.598	0.639
Audio + Text (both)	0.588	0.605	0.610	0.652

Table 5.17: F1 score of online anger for the IEMOCAP dataset ($i = 2, o = 1, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.524	0.553	0.563	0.592
Audio-only, without raw waveform (both)	0.528	0.558	0.567	0.611
Audio-only, with raw waveform (single)	0.533	0.560	0.571	0.614
Audio-only, with raw waveform (both)	0.552	0.562	0.578	0.628
Audio + Text (single)	0.565	0.568	0.586	0.632
Audio + Text (both)	0.576	0.594	0.605	0.643

Table 5.18: F1 score of online anger for the IEMOCAP dataset ($i = 2, o = 2, s = 0$)

Features	DNN	LSTM	Bi-LSTM	EAD
Audio-only, without raw waveform (single)	0.541	0.568	0.585	0.612
Audio-only, without raw waveform (both)	0.543	0.572	0.590	0.617
Audio-only, with raw waveform (single)	0.556	0.576	0.596	0.621
Audio-only, with raw waveform (both)	0.563	0.578	0.604	0.638
Audio + Text (single)	0.577	0.589	0.610	0.642
Audio + Text (both)	0.595	0.616	0.628	0.659

Table 5.19: F1 score of online anger for the IEMOCAP dataset ($i = 3, o = 2, s = 0$)

Moreover, we vary the numbers of input and output utterances similar to the experiments on the Bengali dataset. For all values of i and o , we consider our chosen feature set and fusion combination with EAD as the online anger classifier. The impact of the values of input and output utterances can be found in Table 5.14. Since the values $i = 3$ and $o = 1$ provide the best results, we use these as default values for most of our experiments. Furthermore, for all combinations of the input and output utterances mentioned in Table 5.14, we show the performance of different models and feature sets. The results of the experiments can be found in Tables 5.15, 5.16, 5.17,

5.18, and 5.19. For all values of input and output utterances, our chosen feature set gives the best performance. Similarly, for all cases, EAD performs better than the baseline models.

5.3.3 Discussions and Limitations

We compare our proposed Online Anger Detector (EAD) with different classification schemes for the Bengali call-center dataset and the IEMOCAP dataset. Our approach consistently outperforms the baseline methods in both datasets irrespective of language and domain. Besides, we compare several fusion classifiers and find the best-performing fusion classifier. We consider the utterances of both speakers in a dyadic conversation, which outperforms considering a single speaker discarding the utterances of the other speaker. Besides, we compare the performance of our chosen feature set with different combinations of audio-only feature sets and find that our feature combination performs significantly better than the baselines. We also vary the numbers of input, output, and skipped utterances to find the optimal combination of these values for predicting online anger. We find that increasing the number of input utterances generally improves performance. However, increasing the number of output and skipped utterances naturally reduces the F1 scores of online anger.

There are some limitations to our study. Since we work with real-life spontaneous telephone conversations, there are some missed utterances in the dataset due to noise or both speakers speaking together, affecting the results. In addition, there is a shortage in the amount and variety of training data. The model might perform better with a large amount of training data. Moreover, we perform speech-to-text transcriptions with the help of Google speech-to-text. These transcriptions need highly time-consuming manual corrections. However, in real-life online anger detection, there will be no scope for manual corrections. Therefore, we omit manual corrections in the case of predicting online anger, even though they are used to predict offline anger. Besides, spontaneous conversational Bengali contains regional words, incomplete sentences, wrong pronunciation, and wrong grammatical structure. All of these negatively impact the performance of our model. We keep those errors in our data to make the results consistent with real-time online anger detection since there will be no scope to correct these mistakes. Furthermore, the output data can come from either of the two participants. This adds a high degree of uncertainty to the prediction and does not allow the inclusion of gender information. These reasons explain why the performance of OAD is significantly better than EAD.

Chapter 6

Conclusion

In this research, we have presented the findings of our investigation into offline and online audio-textual anger detection with the help of deep learning ensemble models that incorporate a mid-level fusion of audio and textual features. For offline anger detection, we have proposed the OAD, which adapts an attention-based CNN architecture and incorporates a gender classifier and BERT-based textual features. We have shown that considering the effect of gender can improve the performance of OAD. On the other hand, for online anger detection, we have proposed the EAD, which is a transformer-based ensemble model. We have shown that considering the utterances of both speakers can improve the performance of EAD.

Furthermore, we have compared the performance of our models to several baseline models, feature sets, and fusion combinations using the Bengali call-center dataset and the IEMOCAP dataset. We have observed that our models consistently outperform the baselines in all cases. OAD achieves an F1 score of 91.4% and 85.5% on the IEMOCAP dataset and the Bengali call-center data set, respectively. On the other hand, EAD achieves an F1 score of 67.7% and 66.9% on the IEMOCAP data set and the Bengali call-center data set, respectively. Besides, for EAD, we have varied the utterance parameters to find the parameter combinations that provide the best results for our particular case studies.

In the future, we want to make the anger detection models more versatile by including regional dialects in our work. We believe that this work will provide a significant baseline for future work in audio-textual anger recognition. We also hope that our findings will help shape the integration of anger detection into different personal situations and aid in the development of anger management tools.

References

- [1] F. Abu Shaqra, R. Duwairi, and M. Al-Ayyoub, “Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models,” *Procedia Computer Science*, vol. 151, pp. 37–44, 01 2019.
- [2] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, “Detecting real life anger,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4761–4764, 2009.
- [3] J. Deng, F. Eyben, B. Schuller, and F. Burkhardt, “Deep neural networks for anger detection from real life speech data,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–6, IEEE, 2017.
- [4] J. Devnath, S. Hossain, M. Rahman, H. R. Saha, M. A. Habib, and N. Sultan, “Emotion recognition from isolated Bengali speech,” *Journal of theoretical and applied information technology*, vol. 98, pp. 1523–1533, 2020.
- [5] P. Ekman, “What Scientists Who Study Emotion Agree About,” *Perspectives on Psychological Science*, vol. 11, pp. 31–34, 01 2016.
- [6] A. Khalil, W. Al-Khatib, E.-S. El-Alfy, and L. Cheded, “Anger Detection in Arabic Speech Dialogs,” in *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, pp. 1–6, 2018.
- [7] M. Mertoglu, “Importance of Anger Management in Pre-School Childhood,” *International Journal of Education and Practice*, vol. 6, pp. 200–205, 01 2018.
- [8] N. Alia-Klein, G. Gan, G. Gilam, J. Bezek, A. Bruno, T. Denson, T. Hendler, L. Lowe, V. Mariotti, M. R. A. Muscatello, S. Palumbo, S. Pellegrini, P. Pietrini, and A. Rizzo, “The feeling of anger: From brain networks to linguistic expressions,” *Neuroscience Biobehavioral Reviews*, vol. 108, 12 2019.
- [9] J. M. Lohr, B. O. Olatunji, R. F. Baumeister, and B. J. Bushman, “The psychology of anger venting and empirically supported alternatives that do no harm,” *The Scientific Review*

- of Mental Health Practice: Objective Investigations of Controversial and Unorthodox Claims in Clinical Psychology, Psychiatry, and Social Work*, vol. 5, pp. 53–64, 2007.
- [10] M. Erden and L. M. Arslan, “Automatic detection of anger in human-human call center dialogs,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [11] E. Miron-Spektor and R. Anat, “The effects of anger in the workplace: When, where, and why observing anger enhances or hinders performance,” *Research in Personnel and Human Resources Management*, vol. 28, pp. 153–178, 02 2009.
- [12] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, “Anger Recognition in Speech Using Acoustic and Linguistic Cues,” *Speech Communication*, vol. 53, pp. 1198–1209, 11 2011.
- [13] “The World Factbook.” <https://www.cia.gov/the-world-factbook/countries/world/>, 02 2018. Accessed: 2018-02-21.
- [14] H. M. M. Hasan and M. A. Islam, “Emotion Recognition from Bengali Speech using RNN Modulation-based Categorization,” in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1131–1136, 2020.
- [15] A. Mohanta and U. Sharma, “Bengali speech emotion recognition,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 2812–2814, 2016.
- [16] N. Nomoto, M. Tamoto, H. Masataki, O. Yoshioka, and S. Takahashi, “Anger Recognition in Spoken Dialog Using Linguistic and Para-Linguistic Information,” in *INTERSPEECH*, 2011.
- [17] D. Neiberg and K. Elenius, “Automatic recognition of anger in spontaneous speech,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.
- [18] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, “Anger detection in call center dialogues,” in *2015 6th IEEE international conference on cognitive infocommunications (CogInfoCom)*, pp. 139–144, IEEE, 2015.
- [19] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Attention driven fusion for multi-modal emotion recognition,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3227–3231, IEEE, 2020.
- [20] S. Shahriar and Y. Kim, “Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning,” in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, IEEE, 2019.

- [21] J. Mongkolnavin and W. Saewong, "Prediction of Forthcoming Anger of Customer in Call Center Dialogs," *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 779–783, 2020.
- [22] C. C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1983–1986, 09 2009.
- [23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, IEEE, 2018.
- [24] S. Sultana, M. Iqbal, M. Selim, M. Rashid, and M. Rahman, "Bangla Speech Emotion Recognition and Cross-lingual Study Using Deep CNN and BLSTM Networks," *IEEE Access*, vol. 10, pp. 564–578, 01 2022.
- [25] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.
- [26] "Google Speech-to-text." <https://cloud.google.com/speech-to-text/>. Accessed: 2022-02-25.
- [27] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.
- [28] S. Buus, M. Florentine, and T. Poulsen, "Temporal integration of loudness, loudness discrimination and the form of the loudness function," *The Journal of the Acoustical Society of America*, vol. 101, pp. 669–80, 03 1997.
- [29] S. J. H. Jeans, *Science and Music*. reprinted by Dover, 1938.
- [30] C.-h. Chen, *Signal processing handbook*, vol. 51. CRC Press, 1988.
- [31] "Phonetics and Theory of Speech Production." http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap3.html, 1999. Accessed: 2021-02-19.
- [32] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, pp. 625–636, 01 2010.

- [33] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, C. Xu, and Q. Tian, "HMM-Based Audio Keyword Generation," in *Advances in Multimedia Information Processing - PCM 2004*, pp. 566–574, Springer, 2005.
- [34] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.
- [35] J. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.
- [36] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, pp. 175–185, 1992.
- [37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, p. 144–152, Association for Computing Machinery, 1992.
- [38] Y. Goldberg and M. Elhadad, "splitSVM: fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 237–240, 2008.
- [39] F. Rosenblatt, "PRINCIPLES OF NEURODYNAMICS, PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS," *American Journal of Psychology*, vol. 76, p. 705, 1963.
- [40] O. Abiodun, A. Jantan, O. Omolara, K. Dada, N. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, p. e00938, 11 2018.
- [41] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.
- [42] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," *arXiv preprint arXiv:1609.08409*, 2016.
- [43] H. H. Aghdam and E. J. Heravi, "Guide to convolutional neural networks," *New York, NY: Springer*, vol. 10, no. 978-973, p. 51, 2017.
- [44] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.

- [45] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*, pp. 448–456, PMLR, 2015.
- [46] K. Yamaguchi, K. Sakamoto, T. Akabane, and Y. Fujimoto, “A neural network for speaker-independent isolated word recognition,” in *ICSLP*, 1990.
- [47] M. Lin, Q. Chen, and S. Yan, “Network in network,” *arXiv preprint arXiv:1312.4400*, 2013.
- [48] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” *ArXiv*, vol. abs/1706.03762, 2017.
- [49] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.
- [50] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.
- [51] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [52] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
- [53] M. Lech, M. Stolar, C. Best, and R. Bolia, “Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding,” *Frontiers in Computer Science*, vol. 2, p. 14, 2020.
- [54] Y. Wang and L. Guan, “An investigation of speech-based human emotion recognition,” in *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, pp. 15–18, IEEE, 2004.
- [55] N. Washani and S. Sharma, “Speech Recognition System: A Review,” *International Journal of Computer Applications*, vol. 115, pp. 7–10, 04 2015.
- [56] Q. Al-Shayea and M. Al-Ani, “Speaker Identification: A Novel Fusion Samples Approach,” *International Journal of Computer Science and Information Security*, vol. 14, pp. 423–427, 07 2016.
- [57] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, “A Review of Deep Learning Based Speech Synthesis,” *Applied Sciences*, vol. 9, p. 4050, 09 2019.
- [58] S. Koolagudi and K. Rao, “Emotion recognition from speech: A review,” *International Journal of Speech Technology*, vol. 15, 06 2012.

- [59] U. A. Asiya and V. K. Kiran, "Speech Emotion Recognition-A Deep Learning Approach," in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 867–871, 2021.
- [60] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.
- [61] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 04 2003.
- [62] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of German emotional speech," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.
- [63] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.
- [64] Ç. Oflazoglu and S. Yildirim, "Turkish emotional speech database," in *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, pp. 1153–1156, IEEE, 2011.
- [65] V. Makarova and V. A. Petrushin, "RUSLANA: A database of Russian emotional utterances," in *Seventh international conference on spoken language processing*, 2002.
- [66] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6685–6689, IEEE, 2019.
- [67] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition," in *INTERSPEECH*, pp. 1656–1660, 2019.
- [68] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One Source to Detect them All: Gender, Age, and Emotion Detection from Voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 338–343, 2021.
- [69] C. H. Wu, Z. J. Chuang, and Y. C. Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM transactions on Asian language information processing (TALIP)*, vol. 5, no. 2, pp. 165–183, 2006.
- [70] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni, "Emotion Recognition from Text Based on Automatically Generated Rules," in *2014 IEEE International Conference on Data Mining Workshop*, pp. 383–392, 2014.

- [71] D.-A. Phan, H. Shindo, and Y. Matsumoto, "Multiple emotions detection in conversation transcripts," in *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pp. 85–94, 2016.
- [72] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
- [73] C. Huang, A. Trabelsi, and O. R. Zaïane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," *arXiv preprint arXiv:1904.00132*, 2019.
- [74] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–9, 12 2019.
- [75] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2594–2604, Association for Computational Linguistics, 2018.
- [76] V. Jha, N. Prakash, and S. Sagar, "Wearable anger-monitoring system," *ICT Express*, vol. 4, no. 4, pp. 194–198, 2018.
- [77] D. Waqar, T. Gunawan, M. Morshidi, and M. Kartiwi, "Design of a Speech Anger Recognition System on Arduino Nano 33 BLE Sense," in *2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pp. 64–69, 08 2021.
- [78] A. Azman, K. J. Raman, I. A. J. Mhlanga, S. Z. Ibrahim, S. Yogarayan, M. F. A. Abdullah, S. F. A. Razak, A. H. M. Amin, and K. S. Muthu, "Real Time Driver Anger Detection," in *International Conference on Information Science and Applications 2018 (ICISA 2018)*, pp. 157–167, Springer, 2019.
- [79] I. Nicolaidou, F. Tozzi, and A. Antoniadou, "A gamified app on emotion recognition and anger management for pre-school children," *International Journal of Child-Computer Interaction*, vol. 31, p. 100449, 12 2021.
- [80] L. Yan, P. Wan, and D. Zhu, "The Induction and Detection Method of Angry Driving: Evidences from EEG and Physiological Signals," *Discrete Dynamics in Nature and Society*, vol. 2018, pp. 1–16, 08 2018.

- [81] J. Pohjalainen and P. Alku, "Automatic detection of anger in telephone speech with robust autoregressive modulation filtering," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 05 2013.
- [82] Mohamed Ezzeldin A. Elshaer and Scott Wisdom and Taniya Mishra, "Transfer learning from sound representations for anger detection in speech," *ArXiv*, vol. abs/1902.02120, 2019.
- [83] F. M. Lee, L. Hua, and R. Y. Huang, "Recognizing low/high anger in speech for call centers," *Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA'08)*, 01 2008.
- [84] C. Zheng, C. Wang, and J. Ning, "An Ensemble Model for Multi-Level Speech Emotion Recognition," *Applied Sciences*, vol. 10, p. 205, 12 2019.
- [85] J. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep Learning for Time Series Forecasting: A Survey," *Big Data*, vol. 9, 12 2020.
- [86] O. Sezer, U. Gudelek, and M. Ozbayoglu, "Financial time series forecasting with deep learning : A systematic literature review: 2005–2019," *Applied Soft Computing*, vol. 90, p. 106181, 02 2020.
- [87] A. Tealab, "Time Series Forecasting using Artificial Neural Networks Methodologies: A Systematic Review," *Future Computing and Informatics Journal*, vol. 3, 11 2018.
- [88] Y. Kim and E. Mower Provost, "Emotion spotting: discovering regions of evidence in audio-visual emotion expressions," *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*, pp. 92–99, 10 2016.
- [89] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based Emotion Recognition and Next Reaction Prediction," *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 05 2017.
- [90] R. A. Tuhin, B. Paul, F. Nawrine, and A. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 360–364, 02 2019.
- [91] N. Tripto and M. E. Ali, "Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–6, 09 2018.

- [92] M. Rafi-Ur-Rashid, M. Mahbub, and M. A. Adnan, “Breaking the Curse of Class Imbalance: Bangla Text Classification,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 01 2022.
- [93] S. Azmin and K. Dhar, “Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier,” *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–5, 12 2019.
- [94] M. Asimuzzaman, P. D. Nath, F. Hossain, A. Hossain, and R. M. Rahman, “Sentiment analysis of bangla microblogs using adaptive neuro fuzzy system,” in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1631–1638, 2017.
- [95] S. Abu Taher, K. Afsana Akhter, and K. Azharul Hasan, “N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine,” *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5, 09 2018.
- [96] “OpenSMILE.” <https://audeering.github.io/opensmile/>, 02 2021. Accessed: 2021-02-19.
- [97] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *ArXiv*, vol. abs/1810.04805, 2019.
- [98] A. Bhattacharjee, T. Hasan, K. Samin, M. S. Islam, W. U. Ahmad, A. Iqbal, M. S. Rahman, and R. Shahriyar, “BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla,” *arXiv*, 2021.
- [99] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,” *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 01 2010.
- [100] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators,” *ArXiv*, vol. abs/2003.10555, 2020.
- [101] R. Alkhawaldeh, “DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network,” *Scientific Programming*, vol. 2019, pp. 1–12, 09 2019.
- [102] M. Xu, F. Zhang, and W. Zhang, “Head fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset,” *IEEE Access*, vol. 9, pp. 74539–74549, 2021.

-
- [103] V. Scotti, F. Galati, L. Sbattella, and R. Tedesco, “Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition,” in *International Conference on Pattern Recognition*, pp. 114–128, Springer, 2021.
- [104] M. Shami and W. Verhelst, “An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech,” *Speech communication*, vol. 49, no. 3, pp. 201–212, 2007.
- [105] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive rnn for emotion detection in conversations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, 2019.
- [106] J. Grigsby, Z. Wang, and Y. Qi, “Long-Range Transformers for Dynamic Spatiotemporal Forecasting,” *ArXiv*, vol. abs/2109.12218, 2021.

Generated using Postgraduate Thesis L^AT_EX Template, Version 1.03. Department of
Computer Science and Engineering, Bangladesh University of Engineering and
Technology, Dhaka, Bangladesh.

This thesis was generated on Saturday 14 May, 2022 at 8:31 am.