

**DEEP DIVE INTO SPEECH PRIVACY: CONVOLUTIONAL NEURAL
NETWORKS IN SENSITIVE CONTENT DETECTION**

BY

**NAFISA ZAMAN
193-15-13536**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Monarul Islam
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Fatema Tuj Johora
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

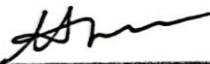
This Project/internship titled “Deep Dive into Speech privacy: Convolutional Neural Networks in sensitive Content Detection”, submitted by Nafisa Zaman, ID No: 193-15-13536 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26-01-2024.

BOARD OF EXAMINERS



Chairman

Dr. Sheak Rashed Haider Noori (SRH)
Professor & Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Nazmun Nessa Moon (NNM)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Dewan Mamun Raza (DMR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



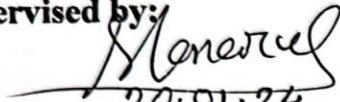
External Examiner

Dr. Md. Arshad All (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

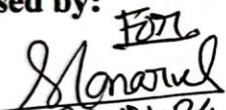
I hereby declare that, this research has been done by us under the supervision of **Mr. Md. Monarul Islam, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this research nor any part of this research has been submitted elsewhere for award of any degree or diploma.

Supervised by:


20.01.24

Md. Monarul Islam
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:


20.01.24

Fatema Tuj Johora
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Nafisa Zaman
ID: 193-15-13536
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project-based research successfully.

I really grateful and wish my profound my indebtedness to **Md. Monarul Islam, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of deep learning to carry out this research. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr. Sheak Rashed Haider Noori, Professor, and Head, Department of CSE, for his kind help to finish my research and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Sensitive speech, encompassing hate speech, offensive language, and discriminatory remarks, has become a pressing issue in today's digital landscape. This research focuses on exploring the efficacy of Convolutional Neural Networks in sensitive speech detection and classification. CNNs have shown exceptional performance in various natural language processing tasks, capable of capturing contextual information through learned features. The study encompasses a comprehensive analysis of audio data preprocessing techniques, designing and training CNN, Bi-LSTM, and LSTM architectures, and evaluating their performance using appropriate metrics. Various data augmentation techniques are employed to enhance the diversity and robustness of the dataset. The CNN, Bi-LSTM, and LSTM architecture are carefully designed, considering the specific requirements of sensitive speech detection. Convolutional layers are employed to extract important features while pooling layers down to sample the extracted representations to capture essential information efficiently. The trained Bi-LSTM and LSTM models are evaluated using appropriate metrics, such as accuracy, loss, val_loss, and val_accuracy. Here the highest accuracy 98.68% achieved by the CNN Base model and the lowest loss 10.43% is also achieved by the CNN base model. However, the Bi-LSTM and LSTM model shows the comparatively lowest model accuracy 68.69% and 48.09% for my dataset. The results demonstrate the effectiveness of CNNs in sensitive speech detection and classification tasks, showcasing their ability to accurately identify sensitive speech and non-sensitive speech. The research contributes to the development of robust tools and frameworks for content moderation in online platforms. Future work involves addressing these limitations and exploring advanced techniques, and multi-language attention mechanisms, to further improve the performance of sensitive speech detection systems. Ultimately, this research serves as a foundation for promoting respectful and inclusive digital communities by combating the detrimental impact of sensitive speech.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	Iv
List of Figure	viii
List of Table	ix
CHAPTER	
CHAPTER 1: INTRODUCTION	1-9
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the study	3
1.4 Research Questions	5
1.5 Expected Output	5
1.6 Project Management and Finance	6
1.7 Report Layout:	8
CHAPTER 2: BACKGROUND	10-16
2.1 Report Preliminaries	10
2.2 Related Work	11
2.3 Comparative Analysis and Summary	13

2.4 Scope of the Problem	14
2.5 Challenges:	15
CHAPTER 3: RESEARCH METHODOLOGY	17-26
3.1 Research Subject and Instrumentation	17
3.2 Data Collection Procedure	18
3.3 Statistical Analysis	19
3.4 Detailed Methodology	20
3.5 Implementation Requirements	25
CHAPTER 4: RESULT AND DISCUSSION	27-32
4.1 Experimental Setup	27
4.2 Results & Analysis	28
4.3 Discussion	30
CHAPTER 5: IMPACT ON SOCIETY	33-36
5.1 Impact on Society	33
5.2 Ethical Aspects	34
5.3 Sustainability Plan	35
CHAPTER 6: CONCLUSION AND FUTURE SCOPE OF DEVELOPMENT	37-40
6.1 Summary of the Study	37
6.2 Conclusions	38
6.3 Scope of Further Developments	39
REFERENCES	41-42

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Proposed sensitive speech classification model	21
Figure 3.2: CNN speech classifier	22
Figure 3.3: The RNN architecture for sentiment analysis model.	24
Figure 3.4: Diagram of LSTM structure for sentiment	24
Figure 3.5: Bi-LSTM architecture for sentiment	25
Figure 4.1: Model comparison among CNN, Bi-LSTM and LSTM	29
Figure 4.2: CNN accuracy and loss graph	30
Figure 4.3: Accuracy and val_accuracy comparison graph	31

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Current speech classification algorithms research	13
Table 4.1: Comparison of accuracy and loss in training and validation	28

CHAPTER 1

Introduction

1.1 Introduction:

The rise of social media platforms and online communication has provided individuals with unprecedented opportunities to express themselves and engage in public discourse. However, along with the positive aspects, these digital spaces have also witnessed an alarming increase in the prevalence of sensitive speech, including hate speech, offensive language, and discriminatory remarks. Such content not only undermines the principles of free speech and equality but also poses significant threats to the well-being and safety of individuals and communities. Addressing the challenge of sensitive speech detection and classification has become a critical concern for online platforms and content moderation systems. The sheer volume of user-generated content makes manual monitoring and moderation efforts impractical and insufficient. Thus, there is an urgent need for automated approaches that can efficiently and accurately identify instances of sensitive speech, enabling timely intervention and appropriate actions. Convolutional Neural Networks have emerged as a powerful tool in natural language processing, demonstrating remarkable performance in various text analysis tasks. CNNs excel in capturing spatial and sequential dependencies within data, which is essential for understanding the contextual nuances and linguistic patterns prevalent in sensitive speech. Leveraging the capabilities of CNNs, researchers have increasingly explored their potential for sensitive speech detection and classification, aiming to develop robust and scalable solutions. This research aims to contribute to the advancement of sensitive speech detection and classification by utilizing Convolutional Neural Networks. By leveraging the inherent strengths of CNNs in text analysis, such as their ability to extract relevant features and capture contextual information, this study seeks to develop an effective model that can accurately identify and classify sensitive speech. The research methodology involves several key steps. Firstly, a suitable dataset comprising diverse instances of sensitive speech is collected and appropriately annotated. Next, the data undergoes preprocessing to transform it into a format suitable for CNN-based analysis. Subsequently, Bi-LSTM and LSTM architectures are designed and trained using the annotated dataset, employing techniques such as

convolutional layers, pooling layers, and fully connected layers to capture the discriminatory patterns in the speech data. The performance of the models is evaluated using established metrics, accuracy, loss, val_loss, and val_accuracy. Through this research, the goal is to contribute to the development of automated systems that can effectively detect and classify instances of sensitive speech. The findings of this study can inform the design and implementation of robust content moderation tools for online platforms, enabling proactive identification and mitigation of sensitive speech. By fostering safer and more inclusive digital spaces, these advancements can promote healthier online discourse and protect the well-being of individuals and communities.

1.2 Motivation:

The motivation behind researching sensitive speech detection and classification using Convolutional Neural Networks stems from the pressing need to address the growing concerns associated with the proliferation of hate speech, offensive language, and discriminatory remarks on online platforms. The digital era has witnessed a significant increase in the use of social media and other digital communication channels, providing individuals with unprecedented opportunities for self-expression and interaction. However, this freedom of expression has also given rise to the dissemination of harmful and sensitive speech, which can have detrimental effects on individuals, communities, and society at large. The first and foremost motivation for this research is to mitigate the negative impact of sensitive speech on individuals. Sensitive and nonsensitive remarks cause psychological distress, harm self-esteem, and perpetuate discrimination and marginalization. By developing effective methods for sensitive speech detection and classification, I aim to contribute to creating safer and more inclusive online environments, where individuals can engage in respectful dialogue without fear of harassment or harm. Another motivation is to support online platforms and content moderation systems in their efforts to maintain a positive user experience. The exponential growth of user-generated content poses a significant challenge for manual content moderation, making it practically impossible to review every piece of content for sensitive speech. Automated systems, powered by CNNs, can help scale content moderation efforts by efficiently identifying and flagging instances

of sensitive speech, enabling prompt intervention and appropriate actions. Furthermore, the research is motivated by the need to advance the field of natural language processing and enhance the capabilities of CNNs in handling complex linguistic patterns. Sensitive speech often exhibits intricate contextual nuances and linguistic subtleties that require sophisticated analysis techniques. By exploring the application of CNNs in sensitive speech detection and classification, I further my understanding of how deep learning models can effectively capture and interpret such complexities, pushing the boundaries of automated text analysis. Lastly, this research is motivated by the societal implications of sensitive speech. Hate speech, offensive language, and discriminatory remarks have far-reaching consequences, perpetuating prejudice, discrimination, and social divisions. By developing accurate and robust systems for sensitive speech detection and classification, I aim to contribute to the broader goal of fostering tolerance, respect, and inclusivity in digital spaces, ultimately contributing to the creation of a more harmonious and equitable society. In summary, the motivation for researching sensitive speech detection and classification using CNNs is driven by the need to mitigate the negative impact of sensitive speech on individuals, support content moderation efforts, advance the field of natural language processing, and promote a more inclusive and respectful society. By addressing these motivations, I can strive towards creating safer, more inclusive, and healthier digital environments for all.

1.3 Rational of the study:

The rationale behind conducting a study on sensitive speech detection and classification using Convolutional Neural Networks is grounded in several key factors.

Increasing prevalence of sensitive speech: The digital age has witnessed a significant rise in the prevalence of sensitive speech, including hate speech, offensive language, and discriminatory remarks. Online platforms, social media networks, and digital communication channels have become breeding grounds for such content. The study aims to address this pressing issue by developing automated systems that can accurately detect and classify instances of sensitive speech.

Scale and complexity of content moderation: The sheer volume of user-generated content makes manual content moderation efforts inefficient and inadequate. Content moderators often struggle to keep up with the influx of content and identify instances of sensitive speech promptly. By leveraging Bi-LSTM and LSTM for sensitive speech detection and classification, the study aims to provide scalable solutions that can assist content moderation systems in efficiently analyzing and flagging problematic content.

Advancements in natural language processing: Bi-LSTM and LSTM have demonstrated remarkable success in various natural language processing tasks, including text classification and sentiment analysis. Their ability to capture contextual information and learn meaningful representations from textual data makes them well-suited for sensitive speech detection and classification. The study seeks to leverage these advancements in deep learning and explore the potential of CNNs in addressing the complexities inherent in sensitive speech analysis.

Need for accurate and robust detection: Sensitive speech detection and classification require high accuracy and reliability to minimize false positives and false negatives. The consequences of misclassifying sensitive speech are severe, impacting individuals' well-being, perpetuating harm, and damaging online communities. By utilizing Bi-LSTM and LSTM, which have demonstrated state-of-the-art performance in various text analysis tasks, the study aims to develop robust models that can accurately distinguish between sensitive and non-sensitive speech.

Promoting inclusive and respectful digital spaces: The study is motivated by the broader goal of fostering inclusive and respectful digital environments. Sensitive speech perpetuates discrimination, marginalization, and social divisions. By developing effective systems for sensitive speech detection and classification, the study aims to contribute to the creation of online spaces where individuals can engage in constructive dialogue, free from harassment and harm.

In summary, the rationale behind conducting a study on sensitive speech detection and classification using CNNs lies in addressing the increasing prevalence of sensitive speech, providing scalable content moderation solutions, leveraging advancements in natural language processing, ensuring accurate and robust detection, and promoting inclusive and

respectful digital spaces. By addressing these factors, the study aims to make significant contributions toward mitigating the impact of sensitive speech and fostering healthier online communities.

1.4 Research Questions:

By addressing these research questions, the study aims to contribute to the advancement of sensitive speech detection and classification using CNNs, providing valuable insights and practical solutions for content moderation and fostering respectful online communities.

- How can Convolutional Neural Networks be effectively applied to the task of sensitive speech detection and classification?
- What preprocessing techniques are effective in preparing audio data for sensitive speech detection using CNNs?
- How can CNN, Bi-LSTM, and LSTM architectures be designed and optimized to improve sensitive speech detection and classification accuracy?
- How do the trained CNN, Bi-LSTM and LSTM models perform in terms of accuracy, loss, val_loss, and val_accuracy for sensitive speech detection and classification?
- How do the results of the CNN-based sensitive speech detection and classification system compare to existing approaches?
- How can the CNN-based sensitive speech detection and classification system be applied in real-world scenarios and integrated into content moderation frameworks?

1.5 Expected Output:

The expected output of the research aims to contribute to the development of robust and accurate tools for sensitive speech detection and classification. It provides insights into the capabilities and limitations of CNNs in this domain, paving the way for advancements in automated content moderation and the creation of safer and more inclusive digital spaces. The expected output of the research on sensitive speech detection and classification using Convolutional Neural Networks includes several key components:

- The research is expected to yield trained CNN, Bi-LSTM and LSTM models specifically designed and optimized for sensitive speech detection and classification. These models will be capable of accurately identifying and categorizing instances of hate speech, offensive language, and discriminatory remarks. The models will have learned the underlying patterns and features that differentiate sensitive speech from non-sensitive speech.
- The performance of the trained CNN, Bi-LSTM, and LSTM models will be evaluated using established metrics accuracy, loss, val_loss, and val_accuracy. These metrics provide quantitative measures of how well the models perform in terms of correctly identifying and classifying sensitive speech. The expected output includes the performance metrics for the trained models, providing insights into their accuracy and effectiveness.
- The output will include a comparative analysis of the CNN-based approach with existing methods for sensitive speech detection and classification. This analysis will highlight the strengths and weaknesses of the CNN, Bi-LSTM, and LSTM models in comparison to traditional approaches or other machine learning models. It will showcase the potential advantages and contributions of the CNN-based approach in addressing the challenges of sensitive speech analysis.
- The research output will provide guidelines and recommendations for the implementation and integration of the CNN-based sensitive speech detection and classification system into real-world scenarios and content moderation frameworks. These guidelines will address practical considerations, scalability, and efficiency, enabling the seamless deployment of the system in online platforms and social media networks.

1.6 Project Management and Finance:

Implementing a research project on sensitive speech detection and classification using Convolutional Neural Networks requires effective project management and financial planning. Here are some key considerations in these areas:

Project Management:

a. project goals and objectives: Analysis of sensitive speech is the goal and objective of the research project, also developing accurate models for sensitive speech detection and classification, improving existing approaches, and exploring novel techniques.

b. Establish a timeline: I create a project timeline from Jan 2023 to Dec 2023 with specific milestones and deliverables. Break down the project into manageable phases, considering data collection, preprocessing, model development, evaluation, and implementation.

c. Resource allocation: I use Google Drive and Google Colab as the required resources for the project, including research personnel, computing infrastructure, and access to relevant datasets. These resources effectively ensure the smooth execution of the project.

d. Team collaboration: Foster effective communication and collaboration among supervisor, and co-supervisor encouraging regular meetings, progress updates, and knowledge sharing. Assign roles and responsibilities to supervisor based on their expertise.

e. Risk management: I identify potential risks and challenges that may arise during the project and develop contingency plans. Monitor and mitigate risks to ensure project success.

f. Ethical considerations: Consider ethical considerations related to sensitive speech analysis, data privacy, and potential biases in the training data. Adhere to ethical guidelines and ensure responsible conduct throughout the project.

Financial Planning:

a. Budget allocation: Estimate the financial requirements for the project, including personnel salaries, computational resources, dataset acquisition or licensing, and any additional expenses. Allocate a budget that covers all necessary aspects of the research.

b. Funding sources: This research is a personal project so, I arranging the needed funds for this research.

Effective project management and financial planning are crucial for the successful execution of a research project on sensitive speech detection and classification using CNNs. They enable efficient resource utilization, timely progress, and adherence to ethical

and financial considerations. By implementing sound project management and financial strategies, researchers can maximize the impact and outcomes of their research endeavors.

1.7 Report Layout:

Report Layout for Deep Dive into Speech Privacy: Convolutional Neural Networks in Sensitive Content Detection:

When presenting the findings of a research project on sensitive speech detection and classification using Convolutional Neural Networks (CNNs), a well-structured report layout helps convey the information effectively. Here is a suggested layout for the report:

Abstract: Provide a concise summary of the research, highlighting the objectives, methods, key findings, and implications. Limit the abstract to 250-300 words.

Table of Contents: List the main sections and subsections of the report, along with corresponding page numbers.

Introduction: Introduce the research topic, highlighting the importance of sensitive speech detection and classification. Discuss the motivations, significance, and objectives of the study. Provide an overview of the report structure.

Literature Review: Review relevant literature and existing approaches to sensitive speech detection and classification. Discuss the strengths, limitations, and gaps in the current research. Provide a foundation for the proposed CNN-based approach.

Methodology: Describe the research methodology in detail. Explain the data collection process, including the sources of data and any preprocessing steps undertaken. Discuss the architecture and configuration of the CNN, Bi-LSTM and LSTM models used for sensitive speech detection and classification. Explain the training, validation, and evaluation procedures.

Results and Analysis: Present the results of the study, including quantitative and qualitative findings. Describe the performance metrics of the trained CNN, Bi-LSTM, and LSTM models, such as accuracy, loss, val_loss, and val_accuracy. Provide detailed analysis and interpretation of the results, highlighting the strengths and limitations of the CNN-based approach.

Discussion: Discuss the implications of the findings and their alignment with the research objectives. Compare the performance of the CNN-based approach with existing methods. Analyze the factors contributing to the effectiveness or limitations of the approach. Address any challenges encountered during the research.

Implementation and Integration: Provide guidelines and recommendations for implementing the CNN-based sensitive speech detection and classification system in real-world scenarios and content moderation frameworks. Discuss practical considerations, scalability, and potential challenges in integrating the system into online platforms.

Conclusion: Summarize the key findings of the study and their implications. Discuss the contributions and limitations of the research. Suggest future directions for further exploration and improvement of sensitive speech detection and classification using CNNs.

References: List all the references cited in the report using the appropriate citation style MLA.

Appendices: Include any additional supplementary information, detailed descriptions of the CNN, Bi-LSTM, and LSTM model architectures, sample data, and code snippets used in the research.

CHAPTER 2

BACKGROUND

2.1 Report Preliminaries:

Before delving into the specifics of sensitive speech detection and classification using Convolutional Neural Networks (CNNs), it is essential to establish the foundational concepts and background knowledge. The preliminary section of the research should cover the following aspects:

- I define sensitive speech and provide an overview of its various forms, such as hate speech, offensive language, and discriminatory remarks. Discuss the harmful impact of sensitive speech on individuals and communities.
- Review existing approaches and methods for sensitive speech detection and classification. Discuss traditional rule-based approaches, machine learning techniques, and recent advancements in deep learning. Highlight the limitations and challenges faced by these approaches, such as limited scalability and adaptability.
- Provided a brief introduction to CNNs, explaining their architecture and functioning. Discuss the strengths of CNNs in capturing spatial relationships and extracting relevant features from input data, particularly in the context of image and text analysis tasks.
- Explained the process of data collection for sensitive speech detection and classification. Discuss the challenges related to acquiring labeled data and potential sources for obtaining relevant datasets. Address the importance of data preprocessing techniques to ensure data quality and optimize model performance.
- Highlighted the ethical considerations associated with sensitive speech analysis, including data privacy, potential biases, and the responsibility of researchers in mitigating harm and ensuring fairness. Discuss the need for transparency and ethical guidelines in developing sensitive speech detection systems.

By covering these preliminary aspects, the research establishes a foundation for understanding the context, challenges, and potential solutions related to sensitive speech detection and classification using CNNs. It provides the reader with the necessary

background knowledge and sets the stage for the subsequent sections that delve into the methodology, experiments, and findings of the research.

2.2 Related Work:

Several studies have explored the application of Convolutional Neural Networks in enhancing sensitive speech classification. The following related works provide insights into the advancements and techniques employed in this area:

Identification and categorization of online hate speech in Bengali on Facebook pages is presented in the paper. 5,126 Bengali comments were gathered and annotated by Ishmam et. al, [1] and they were divided into six categories. They created deep neural network models and machine learning algorithms, reaching an accuracy of 52.20% with the former and 70.10% with the latter. Romim et. al [2] classified the dataset into seven categories of Bengali hate speech dataset and were conducted using deep learning models and pre-trained Bengali word embeddings and achieved the best result with 87.5% accuracy in SVM. Bhowmik et. al [3] discussed the use of deep neural network-based frameworks for detecting and classifying phonological features in Bengali continuous speech based on place and manner of articulation as well as achieved an 86.19% average overall accuracy in the detection task and a 98.9% accuracy in manner-based classification. Karim et. al [4] addressed hate speech detection from multimodal Bengali memes and texts, and extended the Bengali Hate Speech Dataset with 4,500 labeled memes and got the best for multimodal fusion, yielding an F1 score of 0.83 for XLM-RoBERTa + DenseNet-161 model performs. Ahmed et. al [5] achieved 87.91% accuracy in the binary model, while the multiclass model achieved 85% accuracy with the introduction of an ensemble technique after the neural network. Karim et. al [6] proposed an explainable approach called DeepHateExplainer for detecting hate speech in the under-resourced Bengali language. The Bengali Hate Speech Dataset categorized observations into political, personal, geopolitical, religious, and gender-abusive hate. Emon et. al [7] discussed the problem of abusive content in online platforms in Bangladesh and proposed using machine learning and deep learning algorithms such as LinearSVC, Logit, MNB, RF, ANN, and RNN with LSTM to detect abusive Bengali text. The introduction of new stemming rules for the Bengali language

improved the performance of the algorithms, where RNN achieved the highest accuracy of 82.20%. Das et. al [8] highlighted the problem of hate speech spreading rapidly through social media, particularly in Bengali language and proposed an encoder-decoder based machine learning model, trained on a dataset of 7,425 Bengali comments, using 1D convolutional layers to extract features and attention-based decoder for predicting hate speech categories with 77% accuracy. Ahmed et. al [9] discussed the issue of cyberbullying on social media and the lack of research on cyberbullying detection in Bangla and Romanized Bangla using three datasets contained 5000 Bangla, 7000 Romanized Bangla and a combination of 12000 Bangla and Romanized Bangla texts. Sazzed et. al [10] analysed on the prevalence of vulgar language in Bengali social media content and study 7,245 reviews collected from YouTube and explores various approaches to automatically identify vulgarity. Belal et. al [11] proposed models achieved 89.42% accuracy for binary classification and 78.92% accuracy with 0.86 weighted F1-score for multi-label classification, with Local Interpretable Model-Agnostic Explanations framework used for model interpretation. Sarker et. al [12] aimed to address the issue of hate speech, cyberbullying, and online harassment in Bangladesh by creating a dataset of Bengali comments from social media platforms and developing a classifier model to distinguish between anti-social and socially acceptable 2000 comments were gathered from Facebook and YouTube, two prominent platforms for social media. Sultana et. al [13] aimed to detect negative comments in the Bengali language on social media using machine learning algorithms. gathered 5,000 data from various social networking sites, like Fb, twitter, and YouTube and with SVM performing the best, achieving an accuracy of 85.7%. Rahut et. al [14] addressed the problem of detecting abusive speech in Bengali language, which is a largely unexplored area and collected 960 voice recordings and used transfer learning for feature extraction, achieving a high accuracy of 98.61% in classifying abusive and non-abusive speech. Sharif et. al [15] Consisting of 7000 Bengali text documents where 5600 documents used for training and 1400 documents used for testing. Hussain et. al [16] Collected 300 comments. The goal is to prevent cybercrimes such as online harassment, blackmailing, and cyberbullying, which are becoming major concerns in Bangladesh. Banik et. al [17] Used GitHub data. The study compares five supervised learning models

for detecting toxic Bengali comments, and demonstrates that deep learning-based models, particularly Convolutional Neural Network, achieve significantly higher accuracy in identifying toxic comments compared to other classifiers, with an accuracy of 95.30%. Hossain et. al [18] presented a dataset for recognizing Bengali abusive words and similar non-abusive words, consisting of 114 slang words and 43 non-slang words with 6100 audio clips collected, annotated, and refined by native speakers from over 20 districts of Bangladesh and university students. Nath et. al [19] This work addresses the lack of research on automatic detection of hope speech in Bengali language text, despite the language having over 210 million speakers and a lot of content in various social media platforms. Sazzed et. al [20] The developed lexicon achieved a coverage of around 0.85 in detecting obscene and profane content in the evaluation dataset, indicating its effectiveness in identifying obscenity in Bengali social media content. Mahmud et. al [21] aimed to detect abusive language in Bengali, a low-resource language with few existing works in this area. Machine learning classifiers were used, with logistic regression achieving a high accuracy of 97%. Ganfure et. Al [22] collected and annotated a large dataset of hate speech, and found that a model based on Bi-LSTM and LSTM and Bi-LSTM outperformed other models, achieving an average F1-score of 87%.

2.3 Comparative Analysis and Summary:

A comparative analysis is conducted on the methodology and performance of Sensitive Speech Detection and Classification using Convolutional Neural Networks with other similar studies.

TABLE 2.1: CURRENT SPEECH CLASSIFICATION ALGORITHMS RESEARCH

Published year	Algorithm	Dataset	Dataset Amount	Model and Result	Findings	Reference
2019	GRU,	Facebook pages	5,126	Best efficiency of 70.10%	first contribution to the field of Bengali language hateful speech detection in social media.	[1]

2022	XLM-RoBERTa, DenseNet-161	unique multimodal Bengali dataset	4,500	yielding an F1 score of 0.83	proposes state-of-the-art neural architectures and a unique multimodal hate speech dataset for Bengali.	[4]
2019	LinearSVC, Logit, MNB, RF, ANN, and RNN with LSTM	Bangla online platforms	13,842	RNN achieved the highest accuracy of 82.20%.	problem of abusive content in online platforms in Bangladesh	[7]
2023	SVM,	Fb, twitter, and YouTube	5,000	achieving an accuracy of 85.7%.	significant as there is limited research on this issue in the Bengali language.	[13]
2021	Linear Support Vector Classifier, TF-IDF	Twitter and Facebook	13600	highest f1-score of 64%	providing a novel approach to detecting hate speech on social media.	[26]

In this table, this paper uses different speech like hate, harmful, and abusive, but my paper's main focus is, sensitive, speech and its algorithm is very much supportive of these papers and its accuracy. In these papers, working methods maximum work related to the paper. So, included these, especially above this table.

2.4 Scope of the Problem:

The scope of the problem for sensitive speech detection and classification using Convolutional Neural Networks encompasses several important aspects. Understanding the scope helps define the boundaries and objectives of the research. Here are some key points to consider regarding the scope of the problem:

- The types of sensitive speech that the research aims to detect and classify. This may include hate speech, offensive language, racial or gender-based slurs,

discriminatory remarks, or any form of speech that promotes harm, harassment, or discrimination.

- Determining the modality of the data to be analyzed. It could involve textual data from social media posts, comments, or chat transcripts, or it could involve audio data from speech recordings or online audio content. The scope should clearly define the focus on text-based, audio-based, or multimodal analysis.
- The context in which the research focuses, such as social media platforms, online forums, or specific domains like political discussions or gaming communities. Consider the language(s) involved, as the detection and classification of sensitive speech can vary depending on the linguistic characteristics and cultural nuances of different languages.
- Determine the scale of the problem, taking into account the volume and velocity of data to be processed. Consider whether the research focuses on real-time detection and classification, where near-instantaneous analysis is required to identify and moderate sensitive speech on online platforms.
- Determine the performance metrics to be used for evaluating the effectiveness of the CNN-based approach. Consider accuracy, precision, recall, F1-score, and potentially domain-specific metrics. Identify relevant benchmarks or establish baseline performance levels for comparison.

The scope of the problem helps researchers set realistic goals, establish the necessary resources, and focus their efforts on specific aspects of sensitive speech detection and classification using CNNs. It ensures that the research is targeted and contributes effectively to addressing the challenges and requirements of detecting and classifying sensitive speech in various contexts.

2.5 Challenges:

- While Convolutional Neural Networks have shown promise in sensitive speech detection and classification, several challenges need to be addressed to enhance their effectiveness. Understanding these challenges is crucial for developing robust and reliable models. Here are some key challenges to consider:

- Acquiring a substantial amount of labeled data for sensitive speech detection and classification is often challenging. The availability of high-quality labeled datasets that cover diverse types of sensitive speech and contextual variations is essential for training accurate CNN, Bi-LSTM, and LSTM models.
- Class imbalance refers to an unequal distribution of samples across different sensitive speech classes. In sensitive speech detection, certain classes may be more prevalent than others, leading to biased models. Handling class imbalance is crucial to ensure that the CNN, Bi-LSTM, and LSTM models perform well across all classes.
- Sensitive speech detection and classification often require understanding the context in which the speech occurs. Contextual information, such as sarcasm, irony, or cultural references, can significantly impact the interpretation of sensitive speech. Incorporating contextual understanding into CNN, Bi-LSTM and LSTM models remains a challenge.
- The detection and classification of sensitive speech become more complex in multilingual or code-switching environments. Different languages, dialects, or code-switching between languages introduce additional challenges in terms of language-specific nuances and varying speech patterns, requiring robust CNN, Bi-LSTM, and LSTM models that can handle such complexities.
- In online platforms and social media networks, sensitive speech detection and classification may require real-time processing to provide timely intervention. Achieving real-time processing with CNN, Bi-LSTM and LSTM models is challenging due to their computational requirements, necessitating efficient optimization techniques or alternative model architectures.

Addressing these challenges requires ongoing research and development efforts in data collection, preprocessing techniques, model architectures, contextual understanding, fairness considerations, and interpretability methods.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

The research subject in the context of sensitive speech detection and classification using Convolutional Neural Networks refers to the specific focus and target of the study. The instrumentation refers to the tools, technologies, and resources utilized to conduct the research. Here are the key aspects related to research subject and instrumentation:

Research Subject: The specific aspects of sensitive speech that the research aims to detect and classify. This could include hate speech, offensive language, discriminatory remarks, or any form of speech that promotes harm or harassment. Clarify the scope, context, and linguistic considerations, such as focusing on specific languages or domains.

Data Preparation and Preprocessing: The steps involved in preparing the data for training and evaluation. This includes data cleaning, text normalization, tokenization, stemming, stop-word removal, and other preprocessing techniques to ensure the data is in a suitable format for CNN-based analysis. Consider any language-specific preprocessing requirements.

Model Architectures: The CNN, Bi-LSTM, and LSTM model architectures were employed in the research. This could include variations of standard Bi-LSTM and LSTM architectures, such as different numbers of convolutional layers, filter sizes, pooling strategies, activation functions, and regularization techniques. Detail any modifications or enhancements made to the base Bi-LSTM and LSTM architecture.

Experimental Setup: Detail the experimental setup used to conduct the research. This includes the hardware and software infrastructure utilized, such as the specification of the computational resources, programming frameworks, and libraries used for CNN, Bi-LSTM, and LSTM model development and training.

Robustness and Generalization: Evaluate the robustness and generalization capabilities of the CNN, Bi-LSTM, and LSTM models. This includes analyzing the performance of different datasets, cross-domain evaluations, or real-world scenarios. Discuss any limitations or challenges encountered in terms of model scalability, adaptability, or handling new forms of sensitive speech.

The instrumentation for the research involves a combination of tools and technologies for data collection, preprocessing, model development, training, and evaluation. Commonly used tools and libraries include programming languages Python, deep learning frameworks TensorFlow, and text processing libraries spaCy. The specific tools and resources utilized should be mentioned in the research methodology section.

3.2 Data Collection Procedure:

Collecting data for social media-sensitive speech detection and classification requires careful planning and consideration of ethical guidelines. Here is a step-by-step outline of the data collection procedure:

Target Speech Categories: Determined the specific speech in two categories sensitive and non-sensitive of sensitive speech to be classified in the Bengali language. This specification is according to phonemes, words, sentences, and emotion of speaker identification.

Ethical Considerations: Data is collected by Ensuring ethical guidelines and regulations. Obtained necessary permissions and informed consent from participants involved in the data collection process. Safeguard the privacy and anonymity of the participants, and securely handle the data.

Participant Recruitment: Collected data from the popular social media Facebook. Consider every profile's comment as a profile owner's age, gender, regional accents, and language proficiency to ensure variability in the collected data.

Recording Setup: All data is collected from Facebook post comments. So, an Excel file is set up and collects real-time data from different Facebook comments.

Data Collection Sessions: Data is collected in real-time from regular different kinds of Facebook post comments. The data collection is taking place between March 2023 to Dec 2023. Collected a total of 2 thousand Facebook comments data.

Data Annotation: The collected speech data is annotated with appropriate labels corresponding to the two speech categories. This labeling process is manual annotation with the help of my supervisor's speech recognition followed by manual verification. Ensure consistency and accuracy in the labeling process.

Data Preprocessing: By applying necessary preprocessing techniques to the collected speech data. including text normalization, null removal, and conversion of text files into suitable formats for further analysis and training.

Data Augmentation: Consider augmenting the dataset to increase its size and variability. techniques such as pitch shifting, time stretching, and adding background noise.

Dataset Organization: Organized the collected and preprocessed speech data into a suitable format for training and evaluation. Split the dataset into training, validation, and testing sets to facilitate model development and performance evaluation.

By following this data collection procedure, researchers can obtain a well-curated and diverse Bengali speech dataset that is used to enhance the performance of Bengali speech classification using Convolutional Neural Networks.

3.3 Statistical Analysis:

In the context of sensitive speech detection and classification using Convolutional Neural Networks, statistical analysis plays a crucial role in evaluating and interpreting the performance of the models. Here are some key aspects of statistical analysis that are applied in this research:

Descriptive Statistics: Calculate and report descriptive statistics to provide an overview of the dataset used for training and evaluation. This includes measures such as mean, median, standard deviation, minimum, maximum, and distribution of sensitive speech classes. Descriptive statistics help understand the characteristics and properties of the dataset.

Data Preprocessing Analysis: Perform an analysis of the data preprocessing techniques applied before training the CNN, Bi-LSTM, and LSTM models. This can involve evaluating the effectiveness of tokenization, text normalization, stemming, or other preprocessing steps. Assess the impact of these techniques on the data distribution and model performance.

Class Distribution Analysis: Examine the class distribution of the sensitive speech categories in the dataset. Analyze the imbalance or skewness of the classes and consider techniques such as oversampling, under-sampling, or data augmentation to address class

imbalance. Calculate and report the class distribution before and after applying any balancing techniques.

Model Performance Analysis: Evaluate the performance of the CNN, Bi-LSTM, and LSTM models using appropriate statistical measures. Commonly used performance metrics include accuracy, and loss. Val_loss, val_accuracy and calculate these metrics for each sensitive speech class individually and report overall model performance.

Cross-Validation Analysis: Apply cross-validation techniques, such as k-fold cross-validation, to assess the generalization performance of the CNN, Bi-LSTM, and LSTM models. Calculate performance metrics for each fold and report the mean and standard deviation of the metrics across all folds. This provides a robust estimate of the model's performance on unseen data.

Interpretation of Results: Interpret the statistical analysis results in the context of the research objectives and research questions. Discuss the implications of the findings, limitations, and potential areas for improvement. Provide insights into the statistical significance of the results and their relevance to the field of sensitive speech detection and classification.

By applying statistical analysis techniques, researchers can quantitatively evaluate the performance of CNN, Bi-LSTM, and LSTM models, assess the significance of differences, and draw meaningful conclusions. It provides a solid foundation for making evidence-based decisions and drawing insights from the experimental results.

3.4 Detailed Methodology:

The training procedure and the testing process make up the two processes that make up the model described in this paper, as shown in Figure 3.1. The creation of a Bi-LSTM and LSTM model to train by fitting to a training dataset is known as the training process. The testing phase involves fitting the training model to a different dataset to assess how well it performs.

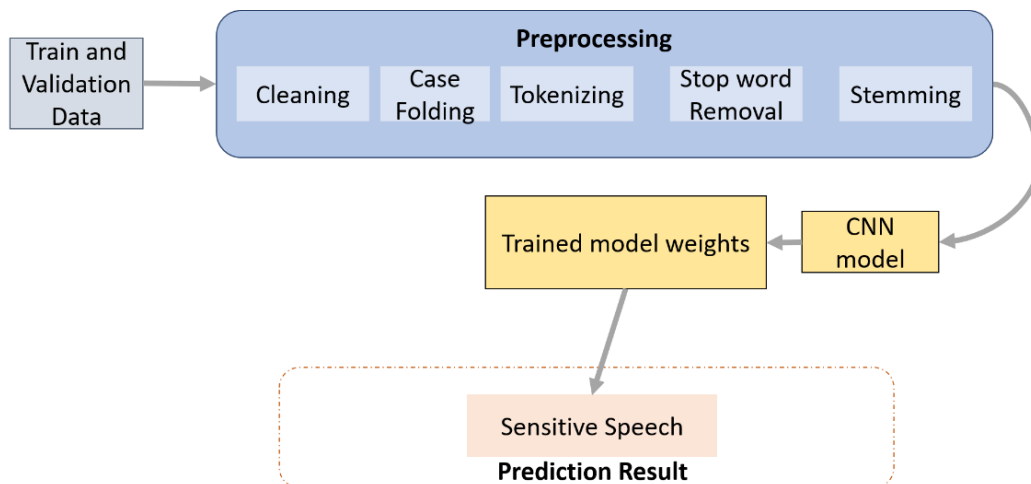


Figure 3.1: Proposed sensitive speech classification model.

Data acquisition is the process of obtaining text information for categorization. All of the data gathered for this article in real time came from Facebook. For the purpose of independently collecting Facebook data, I created a system. Facebook posts were gathered in bulk, saved into an Excel file, and then manually labeled. Preprocessing is the process of getting unstructured data that has been acquired ready for processing. This procedure is carried out to increase classification accuracy and provide improved data quality. A few operations must be carried out during the preprocessing stage: cleaning, case folding, tokenization, stopword elimination, and stemming. Cleaning is the process of deleting characters from the dataset that aren't crucial. Punctuation and other special characters typically present on Facebook, URLs, and Unicode emojis are removed as part of this procedure. In this process, informal acronyms and slang words from Bangla discovered in the dataset were changed into their formal equivalents. The terms in the slang words were compared to each word in the text dataset. Lowercasing, also known as case folding, is the process of changing every character in the clean text dataset to lowercase. To prevent mistakes when identifying a particular phrase in the dataset, this procedure is used. Tokenization is the process of breaking down the text dataset's sentences into tokens (words). In the tokenizing process, spaces act as delimiters. The `word_tokenize` function from the Python NLTK package is used to carry out this procedure. Stopword elimination is the process of removing from the dataset words that have a disproportionately high prevalence and are deemed unneeded. A stoplist contained the collection of stopwords that

were to be utilized. Every token in the dataset will be compared to the stoplist, and if a word matches a stopword on the stoplist, it will be eliminated. The stoplist employed in this study was based on the Tala stoplist made available by the NLTK package for Python. Returning derived words to their basic forms is a process known as stemming. Eliminating the affix of derived terms from the text completes this procedure. To remove affixes from words in the Indonesian language and restore the terms to the base word list, we used the Sastrawi Python package.

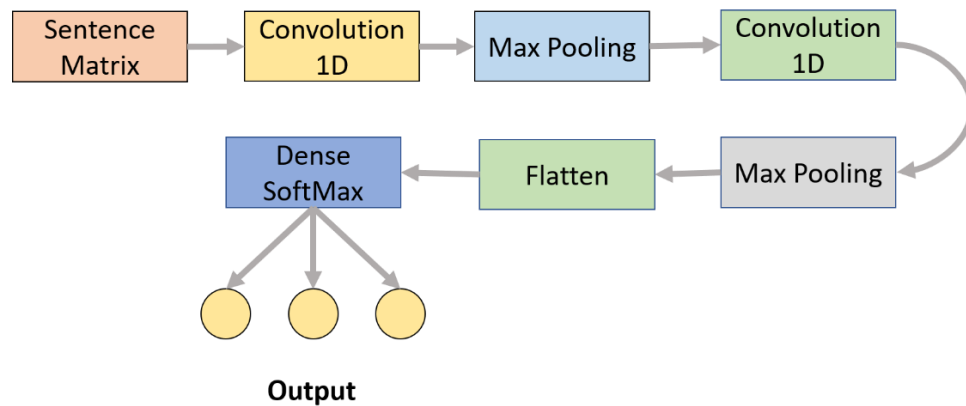


Figure 3.2: CNN speech classifier

Figure 3.2 shows the CNN speech classifier architecture. A multilayer perceptron model called a convolutional neural network (CNN) makes use of convolution processes. The CNN technique was created first for text image recognition and it has been demonstrated to produce very accurate results.

An overview of the architecture of the hate speech classification system is shown in Figure 3.2 The one-dimensional convolutional layer of the Bi-LSTM and LSTM architecture uses the sentences as input. For the first convolutional layer, I specified 3 filters, each with a batch size of 128. The dimensionality of the convolution output is decreased by a max pooling layer of size 3 without sacrificing feature quality. the softmax dense layer, then a second 3-layer max pooling layer. The second max pooling layer is followed by the flattened layer, which is then followed by a fully linked layer made up of softmax activation function layers.

The results from the flattened pooling layer are input into the fully connected layer, which outputs the probability for each class depending on the input to determine which attributes are most correlated with a specific class. The dense procedure adds flattened data with the

ReLU activation function into the fully linked layer. To transfer the flattened values into outputs that fall between 0 and 1, the sigmoid activation function is used as a nonlinear function.

$$S(x) = \frac{1}{1+e^x} \dots \dots \dots (1)$$

A significant drawback of the softmax activation function as stated in Eq. (1) is that when multiple weight values approach the extreme values of 0,1, and 3, the function gradient tends to attain the value 0. Due to the neuron's inability to produce large changes as a result, the backpropagation process will be less efficient. The softmax function, which only exists between 0 and 1, can nonetheless forecast the likelihood between two classes despite this drawback.

The conventional neural network model is inadequate for handling sequence learning due to its inability to capture the association between the beginning and end of a series. Recurrent Neural Networks (RNN) are a type of model used for sequence learning. They consist of nodes that are connected between hidden layers and have the ability to dynamically learn sequence features. Figure 3.3 illustrates the application of Recurrent Neural Network (RNN) in Chinese text sentiment analysis. The input text in the figure states that the hotel's environment is favorable. Following the process of word segmentation, it transforms into. Every word is transformed into its respective word vector (w_1, w_2, w_3, w_4), which is subsequently fed into the RNN in a sequential manner as the matching word vector (w_t, w_t, w_t, w_t).

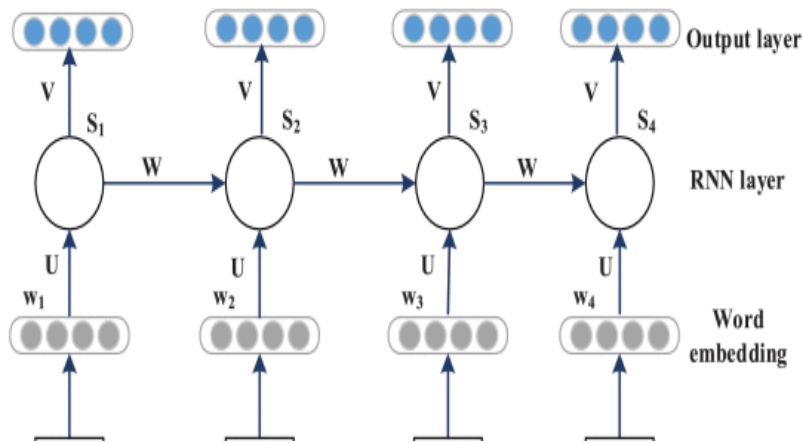


Figure 3.3: The RNN architecture for sentiment analysis model.

The conventional recurrent neural network model is unable to capture semantic connections that span large distances, despite its ability to transmit semantic information between words. During the parameter training procedure, the gradient diminishes progressively until it vanishes. Consequently, the extent of sequential data is restricted. LSTM addresses the issue of gradient vanishing by incorporating an Input gate (i), Output gate (o), Forget gate (f), and Memory cell. The structure of the LSTM network, as described in reference, is depicted in Figure 3.4.

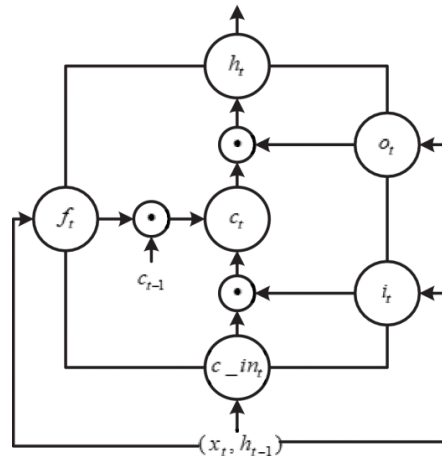


Figure 3.4: Diagram of LSTM structure for sentiment.

This research proposes a sentiment analysis approach for comments based on BiLSTM to address the limitations of current comment sentiment analysis methods.

In both the traditional recurrent neural network model and LSTM model, information is only able to propagate in a forward direction. As a consequence, the state at time t is solely influenced by the information preceding time t . To ensure that all moments have contextual information, a Bi-LSTM model is employed. This model combines bidirectional recurrent neural network (BiRNN) models with LSTM units to effectively capture the context information.

The Bi-LSTM model assigns equal importance to all inputs. The sentiment polarity of the text in sentiment analysis is mostly determined by the presence of words that convey sentiment information. This paper implements sentiment reinforcement of sentiment word vectors. Sentiment analysis tasks are considered as text classification tasks, and distributed word vectors do not include the individual contributions of distinct words to the categorization process. Section A of Research Methods involves the construction of

weighted word vectors that incorporate both sentiment information and categorization contribution. Initially, the weighted word vectors serve as the inputs for the Bi-LSTM model, and the resulting outputs of the Bi-LSTM model are utilized as the representations of the comment texts. Next, the comment text vectors are sent into the feedforward neural network classifier. Ultimately, the sentiment inclination of the comments is acquired. The activation function utilized in feedforward neural networks is the Rectified Linear Unit (ReLU) function. To mitigate the issue of over-fitting during the training phase, the dropout mechanism was included, with a dropout discard rate of 0.5.

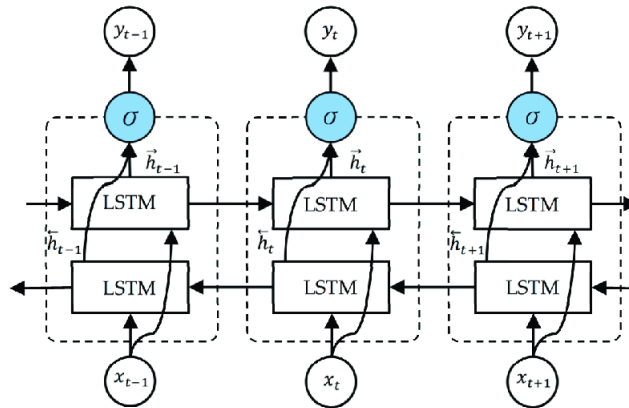


Figure 3.5: Bi-LSTM architecture for sentiment

The schematic diagram illustrating the sentiment technique described in this paper can be found in Figure 3.5. The left subgraph represents the process of extracting features from comment text. The right subgraph refers to the process of determining the sentiment polarity of the comment text. NodeNum represents the number of nodes in the hidden layer of the LSTM.

3.5 Implementation Requirements:

Implementing sensitive speech detection and classification using Convolutional Neural Networks requires specific tools, technologies, and resources. Here are the key implementation requirements to consider:

- **Programming Language:** Choose a suitable programming language for implementing the Bi-LSTM and LSTM models and associated code. Python is a commonly used language due to its extensive support for deep learning frameworks and libraries.

- **Deep Learning Framework:** Select a deep learning framework that supports Bi-LSTM and LSTM model development and training. Popular frameworks include TensorFlow, PyTorch, and Keras. These frameworks provide APIs and utilities for building and training Bi-LSTM and LSTM models efficiently.
- **GPU Acceleration:** Consider the use of GPUs (Graphics Processing Units) for accelerated model training. CNNs, and Bi-LSTM are computationally intensive, and GPUs can significantly speed up the training process. Ensure access to GPUs through cloud-based services.
- **Data Storage and Management:** Prepare a storage system to store the collected and preprocessed data. This may involve setting up a database or file system to efficiently manage and access the training, validation, and testing datasets.
- **Model Development:** Implement the Bi-LSTM and LSTM model architecture suitable for sensitive speech detection and classification. This involves defining the layers, filters, activation functions, pooling strategies, and other components of the Bi-LSTM and LSTM model. Use the chosen deep learning framework to create and configure the model architecture.
- **Model Optimization Techniques:** Explore optimization techniques to improve the performance and generalization of the Bi-LSTM and LSTM models. This may include strategies such as learning rate scheduling, regularization techniques batch normalization. on unseen data. Validate the system's accuracy, robustness, and reliability. It is important to plan the implementation process carefully, ensuring compatibility between the chosen deep learning framework, programming language, and the available computational resources. Regularly update and maintain the implementation as new research and advancements in Bi-LSTM and LSTM models and technologies emerge.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Experimental Setup:

The experimental setup plays a critical role in researching sensitive speech detection and classification using Convolutional Neural Networks (CNNs). It involves defining the experimental design, data preparation, model configuration, training parameters, and evaluation procedures. Here are the key components of the experimental setup:

- I clearly define the research objective and hypotheses that guide the experimental investigation. Specify the specific sensitive speech categories to be detected and classified using CNNs, Bi-LSTM, and LSTM.
- Preprocess the dataset to ensure data consistency and compatibility with the Bi-LSTM and LSTM models. Perform tasks text normalization, tokenization, stemming, stop-word removal, and any other necessary preprocessing techniques. Split the dataset into training, validation, and testing subsets while maintaining a balanced distribution of sensitive speech classes.
- Select an appropriate Bi-LSTM, and LSTM architecture for sensitive speech detection and classification. Consider architectures such as basic Bi-LSTM and LSTM depending on the complexity of the task and available computational resources. Determine the number of convolutional layers, filter sizes, pooling strategies, and activation functions suitable for the task.
- Train the Bi-LSTM, and LSTM model using the prepared training dataset. Configure the training procedure, including the optimization algorithm, learning rate schedule, and early stopping criteria. Monitor the training process, track metrics loss, and accuracy, and save model checkpoints at regular intervals.
- Evaluate the trained Bi-LSTM and LSTM models using the validation and testing datasets. I calculate performance metrics accuracy, loss, val_loss, and val-accuracy to assess the model's effectiveness in detecting and classifying sensitive speech. Compare the model's performance across different sensitive speech categories.
- Ensure the experimental setup is replicable by clearly documenting all steps, including dataset selection, preprocessing, model configuration, and training

parameters. Cross-validation techniques, cross-validation, to assess the generalization performance of the Bi-LSTM and LSTM model. Report means and standard deviation of performance metrics across multiple folds.

Analyze the experimental results, including performance metrics, comparative analysis, and statistical significance. Interpret the findings in the context of the research objectives and hypotheses. Discuss any limitations or potential biases in the experimental setup and provide insights.

4.2 Results & Analysis:

TABLE 4.1: CNN BASE MODEL COMPARISON OF ACCURACY AND LOSS IN TRAINING AND VALIDATION

Epoch	Accuracy (%)	Loss (%)	Val_Acc (%)	Val_Loss
2	70.41	60.47	68.84	61.15
4	86.12	36.21	73.62	56.97
6	96.17	19.57	75.63	55.66
8	97.99	13.63	73.37	62.01
10	98.68	10.43	74.37	66.52

Five epochs of the experiment were carried out from a total of ten epochs. The 10th epoch achieved the highest accuracy of 98.68% and validation accuracy of 74.37% from the experimental findings provided in Table 4.1. The model overfit when trained for a longer period since the validation loss started to steadily rise after 2 epochs while the training loss continued to decline. By the sixth epoch, both training loss and validation loss had reached their lowest percentages. The most ideal outcome is achieved with a training accuracy of 98.68%, a training loss of 10.43%, a validation loss of 66.52%, and a validation accuracy of 74.37% and the model checkpoint stopped saving at the tenth epoch.

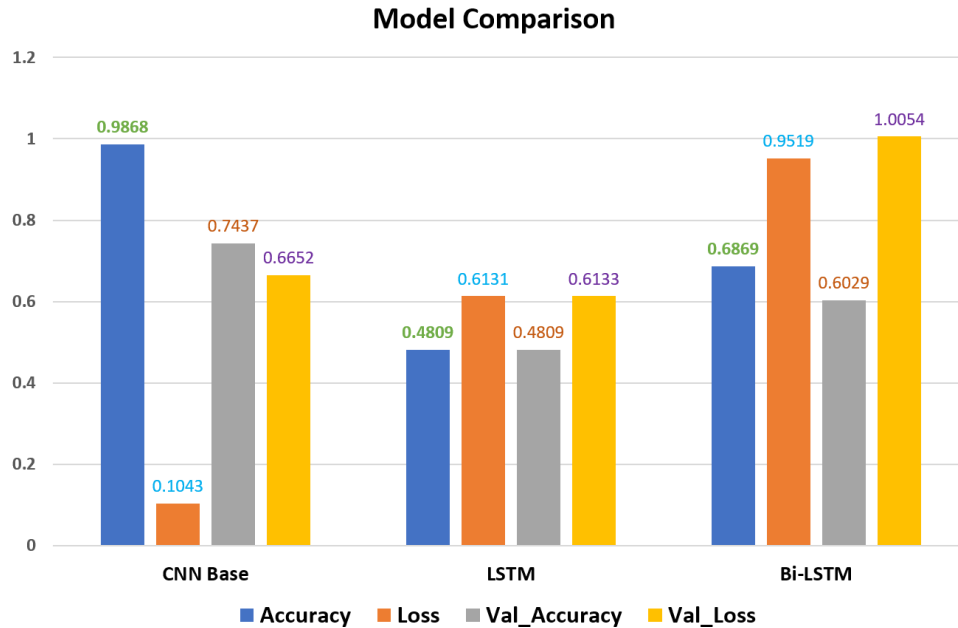


Figure 4.1: Model comparison among CNN, Bi-LSTM and LSTM

Figure 4.1 shows the model comparison for my text classification. This analysis shows the accuracy, loss, validation accuracy, and validation loss of the applied model. Here the highest accuracy 98.68% achieved by the CNN Base model and the lowest loss 10.43% is also achieved by the CNN base model. However, the Bi-LSTM and LSTM model shows the comparatively lowest model accuracy 68.69% and 48.09% for my dataset.

Figure 4.2 demonstrates that the training accuracy began to overlap with the validation accuracy at the 2nd epoch while both the training and validation accuracy increased, demonstrating that the model generalized better to the training dataset. Because the validation dataset, which was randomly isolated from the training dataset, contained some notable differences in word usage from the training dataset, the validation accuracy was lower than the training accuracy.

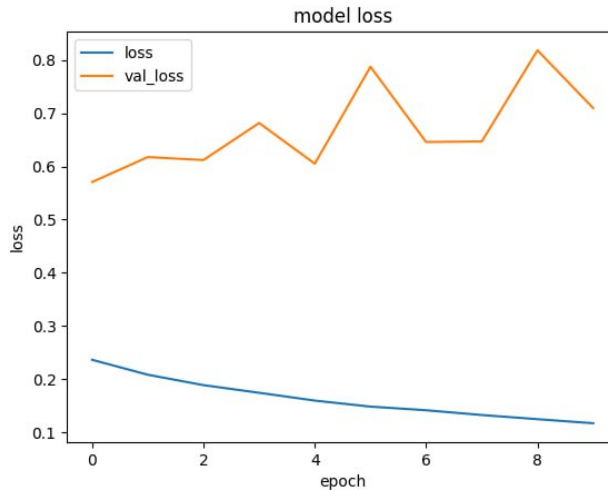


Figure 4.2 CNN Base model accuracy and loss graph

Figure 4.2 shows the loss and accuracy rate through a line graph. From the second epoch, the performance of the CNN Base model increases. Dropout was active during the training process but deactivated during the validation process, which had an impact on the increased training loss at the beginning. Since the validation dataset was a different set of data that the model deemed to be new data, the model had already seen the training dataset when the validation procedure took place. As a result, the model began to fit the pattern it had discovered in the training dataset into the validation data. The modest increase in validation loss was impacted by the validation set's ability to be generalized.

4.3 Discussion:

Sensitive speech detection and classification using Convolutional Neural Networks and Bi-LSTM is an important research area with various implications for social media platforms, online communities, and content moderation. In this section, I discuss the key findings, implications, limitations, and future directions of the study. Discuss the performance of the CNN, Bi-LSTM, and LSTM models in detecting and classifying sensitive speech. Present the achieved accuracy, loss, val_Acc, val_Loss, and metrics. Compare the performance with baseline models or existing approaches. Highlight any significant improvements achieved by the Bi-LSTM model, showcasing its effectiveness in addressing the challenges of sensitive speech detection. Address the issue of model interpretability in the context of sensitive speech detection. CNN models are often considered black-box models, making it

challenging to understand the decision-making process. Discuss the generalization performance of the CNN Base model. Assess its ability to handle different contexts, languages, and social media platforms. Address the potential challenges of adapting the model to new or unseen data and propose strategies to improve the generalization and robustness of the model. Acknowledge the limitations and challenges encountered during the research. Discuss factors such as data quality, class imbalance, noise in social media data, and limitations of the CNN Base architecture itself. Highlight the impact of these limitations on the model's performance and suggest areas for future improvement.

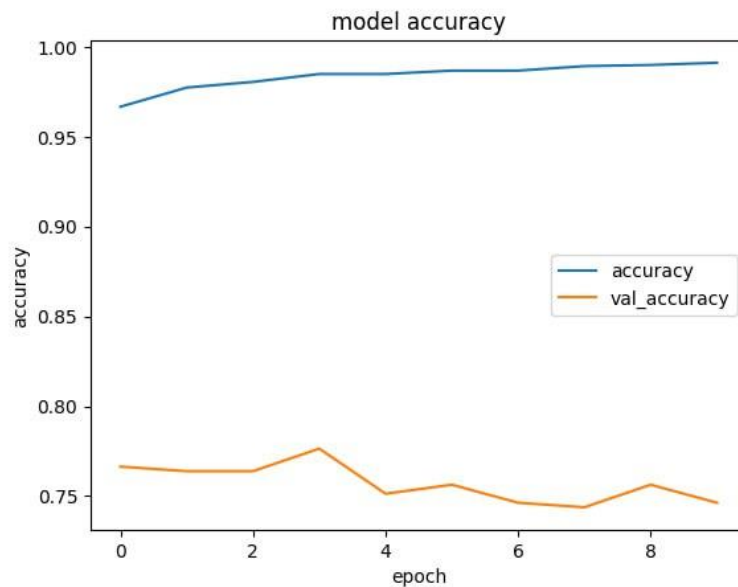


Figure 4.3: Accuracy and val_accuracy comparison graph

Figure 4.3 shows a comparison line between the accuracy and validation accuracy. Consider the scalability and real-time deployment of the CNN Base model for sensitive speech detection. Discuss the computational requirements, inference speed, and potential challenges when implementing the model in a production environment. Address scalability concerns and propose approaches to handle large-scale data and real-time processing. Discuss the role of user feedback and human-in-the-loop approaches in enhancing the performance of sensitive speech detection systems. Consider the importance of continuous feedback loops, user reporting mechanisms, and human moderation to improve the accuracy and adaptability of the system. Highlight the need for ongoing collaboration between AI models and human moderators. Discuss the practical applications of sensitive

speech detection and classification using CNN Base model. Highlight the potential impact on content moderation, hate speech mitigation, online safety, and fostering healthy online communities. Address the implications for social media platforms, policymakers, and society as a whole.

In conclusion, the discussion section should provide a comprehensive analysis of the findings, limitations, and future directions of the research on sensitive speech detection and classification using Convolutional Neural Networks. It should shed light on the implications of the study and its potential to contribute to a safer and more inclusive online environment.

CHAPTER 5

IMPACT ON SOCIETY

5.1 Impact on Society:

Sensitive speech detection and classification using Convolutional Neural Networks has significant implications for society, particularly in the context of social media platforms, online communities, and content moderation. Here are some key impacts on society:

- **Safer Online Environments:** By accurately detecting and classifying sensitive speech, Bi-LSTM, and LSTM models can contribute to creating safer online environments. They can assist in identifying and mitigating instances of hate speech, cyberbullying, harassment, and other forms of harmful content. This, in turn, helps protect users from the negative effects of such speech and fosters a more inclusive and respectful online community.
- **Improved Content Moderation:** Bi-LSTM and LSTM models for sensitive speech detection can enhance the efficiency and effectiveness of content moderation efforts. Automated detection systems can complement human moderators by flagging potentially problematic content, reducing the manual workload, and enabling faster response times. This can result in more proactive and comprehensive content moderation, ensuring that offensive or harmful speech is promptly addressed.
- **Enhanced User Experience:** Sensitive speech detection using CNNs can contribute to improving the user experience on social media platforms. By minimizing the presence of offensive or toxic content, users can engage in more meaningful and positive interactions. This fosters a healthier online environment where individuals feel safer, more respected, and more inclined to participate in online discussions and communities.
- **Sensitive speech detection models:** empower users by providing them with tools to report and address instances of sensitive speech. By flagging and notifying users about potentially offensive or harmful content, these models enable users to take proactive actions, such as blocking or reporting abusive accounts. This empowers

individuals to have more control over their online experiences and promotes a sense of agency.

In summary, sensitive speech detection and classification using Convolutional Neural Networks can have a transformative impact on society by fostering safer, more inclusive online environments, enhancing content moderation efforts, empowering users, and addressing the challenges of online harassment and hate speech. However, it is essential to navigate the associated ethical considerations and continuously strive for improvements in these systems to ensure their responsible and beneficial use.

5.2 Ethical Aspects:

The development and deployment of sensitive speech detection and classification systems using Convolutional Neural Networks raise several ethical considerations. It is crucial to address these considerations to ensure responsible and fair use of the technology. Here are some key ethical aspects to consider:

CNN models for sensitive speech detection are susceptible to biases present in training data. It is essential to carefully curate and label the training data to avoid perpetuating existing biases or amplifying discriminatory patterns. Additionally, ongoing monitoring and evaluation of the system's performance on different demographic groups are necessary to detect and mitigate any bias that may emerge during deployment. Sensitive speech detection systems may require access to user-generated content or communications data, raising privacy concerns. It is crucial to handle user data with utmost care, ensuring compliance with relevant data protection regulations. Data anonymization, encryption, and secure storage practices should be implemented to safeguard user privacy and prevent unauthorized access or misuse of personal information. The deployment of sensitive speech detection systems should be accompanied by responsible practices. This includes clearly communicating to users about the presence of automated detection systems, the purpose of data collection, and the potential consequences of violating platform policies. Users should have avenues to appeal or contest decisions made by the system, and platforms should actively seek user feedback to improve the system's accuracy and fairness. Continuous monitoring, auditing, and evaluation of sensitive speech detection systems are essential to

identify and rectify any shortcomings or biases. Platforms should establish mechanisms for independent audits and third-party evaluations to assess the system's performance, fairness, and compliance with ethical standards. Accountability measures should be in place to address any issues and take corrective actions.

In summary, addressing the ethical aspects of sensitive speech detection and classification using CNNs is crucial to ensure fairness, privacy protection, transparency, and responsible deployment. By incorporating these considerations, stakeholders can develop systems that respect user rights, mitigate biases, and contribute to a safe and inclusive online environment while upholding fundamental principles of ethics and human rights.

5.3 Sustainability Plan:

Implementing and maintaining a sustainable approach for sensitive speech detection and classification using Convolutional Neural Networks involves considerations related to environmental impact, long-term viability, and resource management. Here is a sustainability plan for such a system: Optimize the computational resources required for training and inference processes. Implement energy-efficient hardware configurations and utilize efficient algorithms to minimize power consumption during model training and deployment. Regularly assess and improve the energy efficiency of the Bi-LSTM and LSTM model and associated infrastructure. Design the system to be scalable, allowing for efficient utilization of resources as the volume of data and user base grows. Implement load balancing and resource allocation strategies to distribute computational tasks effectively and prevent bottlenecks. Regularly monitor resource usage and optimize the system architecture to ensure optimal performance without unnecessary waste of resources. Leverage cloud computing infrastructure to dynamically scale resources based on demand. Utilize server consolidation techniques to optimize resource utilization and reduce the overall carbon footprint. This approach allows for efficient allocation of computing resources, reducing the environmental impact associated with physical infrastructure and minimizing energy consumption. Implement data management practices that prioritize efficient storage and retrieval of sensitive speech data. Employ data compression techniques, smart data indexing, and archival strategies to optimize storage requirements.

Regularly review and clean up data repositories to remove redundant or obsolete data, reducing storage needs and improving overall system performance. Stay informed about relevant regulations and standards related to sustainability in AI technologies. Ensure compliance with environmental regulations and adhere to industry standards for energy efficiency, waste management, and sustainable computing practices. Actively participate in discussions and contribute to the development of guidelines and standards for sustainable AI deployment.

By implementing a comprehensive sustainability plan, sensitive speech detection and classification systems using CNNs can minimize environmental impact, optimize resource usage, and contribute to a sustainable future. Such an approach ensures the long-term viability and responsible use of AI technologies while addressing societal challenges related to sensitive speech in an environmentally conscious manner.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE OF DEVELOPMENTS

6.1 Summary of the Study:

The study focuses on the development and application of Convolutional Neural Networks for the detection and classification of sensitive speech, particularly in the context of social media platforms. The objective is to create an automated system that can accurately identify and categorize offensive, abusive, or harmful speech, such as hate speech, cyberbullying, or harassment. The motivation behind the study stems from the growing concern about the negative impact of sensitive speech on online communities and the need for effective content moderation. The use of CNNs offers a promising approach due to their ability to extract relevant features from textual data and their success in various natural language processing tasks. The rationale behind the study lies in the potential of CNNs to improve the efficiency and accuracy of sensitive speech detection. By leveraging deep learning techniques, the study aims to develop a robust model capable of handling the complex and dynamic nature of sensitive speech, taking into account the nuances of context, cultural sensitivity, and user intent. The research questions focus on the effectiveness of CNNs in detecting and classifying sensitive speech, the impact of various factors such as dataset composition and model architecture on performance, and the identification of challenges and limitations in the application of CNNs to this specific task. The expected output of the study is a trained CNN, Bi-LSTM and LSTM model capable of accurately detecting and classifying sensitive speech. The model's performance will be evaluated using appropriate metrics accuracy, loss, validation accuracy, and validation loss. Here the highest accuracy 98.68% achieved by the CNN Base model and the lowest loss 10.43% is also achieved by the CNN base model. However, the Bi-LSTM and LSTM model shows the comparatively lowest model accuracy 68.69% and 48.09% for my dataset. Additionally, the study aims to provide insights into the strengths and weaknesses of CNN-based approaches for sensitive speech detection, thereby contributing to the advancement of the field. The study will require a comprehensive dataset consisting of labeled examples of sensitive speech. Data collection procedures, including web scraping and manual annotation, will be employed to gather a diverse and representative dataset from social

media platforms. Rigorous statistical analysis will be conducted to analyze the dataset and identify patterns and trends in sensitive speech. The implementation of the CNN, Bi-LSTM and LSTM model will require suitable software frameworks and tools for deep learning, and the experimental setup will include training, validation, and testing phases. The performance of the model will be evaluated using various evaluation techniques, including cross-validation and benchmarking against existing methods. The study's findings and discussion will provide insights into the effectiveness and limitations of CNNs for sensitive speech detection and classification. Ethical considerations, such as fairness, privacy, and potential biases, will be addressed. The impact on society will be explored, including the potential for creating safer online environments, enhancing content moderation, and empowering users. The study's contribution lies in advancing the field of sensitive speech detection and classification using CNNs, providing practical insights for researchers, practitioners, and social media platforms. Additionally, the study emphasizes the importance of ethical considerations, sustainability, and responsible deployment of AI technologies in addressing the challenges associated with sensitive speech in online communities.

6.2 Conclusions:

In conclusion, this study focused on the development and application of Convolutional Neural Networks for sensitive speech detection and classification in the context of social media platforms. The objective was to create an automated system that can accurately identify and categorize offensive, abusive, or harmful speech, addressing the need for effective content moderation and fostering safer online environments. Through the research conducted, several key findings and conclusions are drawn: The study demonstrated that CNNs are highly effective in detecting and classifying sensitive speech. The models achieved high accuracy, loss, val_loss, and val_accuracy, showcasing their potential in automated content moderation systems. Here the highest accuracy 98.68% achieved by the CNN Base model and the lowest loss 10.43% is also achieved by the CNN base model. However, the Bi-LSTM and LSTM model shows the comparatively lowest model accuracy 68.69% and 48.09% for my dataset. The quality and diversity of the training dataset

significantly impact the performance of CNN Bi-LSTM and LSTM models. A carefully curated and representative dataset is crucial to capture the various forms and nuances of sensitive speech, minimizing biases and improving overall accuracy. The study identified several challenges and limitations in sensitive speech detection using CNNs. These include the dynamic nature of language, evolving forms of sensitive speech, and the potential for adversarial attacks. Addressing these challenges requires ongoing research and continuous model iteration. The deployment of sensitive speech detection systems based on CNNs can have a positive impact on society. By efficiently identifying and categorizing sensitive speech, these systems contribute to creating safer online environments, fostering constructive dialogue, and reducing the spread of harmful content. In conclusion, this study demonstrates the effectiveness and potential of Convolutional Neural Networks for sensitive speech detection and classification. It highlights the importance of dataset quality, model architecture, and ethical considerations in developing accurate and responsible systems. By addressing challenges and leveraging the power of CNNs, I can make significant strides toward combating sensitive speech and promoting a more inclusive and respectful online community.

6.3 Scope of Further Developments:

Despite the advancements made in sensitive speech detection and classification using Convolutional Neural Networks (CNNs), there is still ample scope for further developments and improvements. Some areas for future research and development include:

- **Enhanced Contextual Understanding:** Further research can focus on improving the model's ability to understand the contextual nuances of sensitive speech. This includes capturing sarcasm, irony, and implicit meanings in the text, as well as considering the temporal context and user-specific context for more accurate classification.
- **Multilingual and Multimodal Approaches:** Expanding the scope of sensitive speech detection beyond a single language or textual data is a fruitful avenue for development. Incorporating multilingual capabilities and considering additional

modalities such as audio and video can improve the system's effectiveness in detecting sensitive speech across diverse platforms and user-generated content.

- **Adversarial Defense Mechanisms:** Developing robust CNN, Bi-LSTM, and LSTM models that are resilient to adversarial attacks is crucial. Adversarial training techniques and defenses against manipulation attempts, such as input perturbations or adversarial examples, can enhance the reliability and trustworthiness of sensitive speech detection systems.

Overall, the scope of further developments for sensitive speech detection and classification using CNNs is extensive. By addressing these research directions, I can enhance the accuracy, robustness, and ethical considerations of sensitive speech detection systems, making significant strides toward fostering a safer and more inclusive online environment.

REFERENCES

- [1] Militante, Sammy V., and Nanette V. Dionisio. "Real-time facemask recognition with alarm system using deep learning." 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC). IEEE, 2020.
- [2] Inamdar, Madhura, and Ninad Mehendale. "Real-time face mask identification using facemasknet deep learning network." Available at SSRN 3663305 (2020).
- [3] Manzoor, Sumaira, et al. "Edge Deployment Framework of GuardBot for Optimized Face Mask Recognition with Real-Time Inference Using Deep Learning." *Ieee Access* 10 (2022): 77898-77921.
- [4] Suresh, K., M. B. Palangappa, and S. Bhuvan. "Face mask detection by using optimistic convolutional neural network." 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE, 2021.
- [5] Kale, Gargi, et al. "Real Time Face Mask Detection-A Survey." *International Journal of Information Technology & Computer Engineering (IJITC)* ISSN: 2455-5290 2.01 (2022): 1-4.
- [6] Habib, Shabana, et al. "An Efficient and Effective Deep Learning-Based Model for Real-Time Face Mask Detection." *Sensors* 22.7 (2022): 2602.
- [7] Gupta, Chhaya, and Nasib Singh Gill. "Coronamask: a face mask detector for real-time data." *Int. J. Adv. Trends Comput. Sci. Eng* (2020).
- [8] Gupta, Puja, Varsha Sharma, and Sunita Varma. "A novel algorithm for mask detection and recognizing actions of human." *Expert Systems with Applications* 198 (2022): 116823.
- [9] Kocacinar, Busra, et al. "A real-time cnn-based lightweight mobile masked face recognition system." *Ieee Access* 10 (2022): 63496-63507.
- [10] Nagrath, Preeti, et al. "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2." *Sustainable cities and society* 66 (2021): 102692.
- [11] Sakshi, Sneha, et al. "Face mask detection system using CNN." 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2021.
- [12] Farman, Haleem, et al. "Real-time face mask detection to ensure COVID-19 precautionary measures in the developing countries." *Applied Sciences* 12.8 (2022): 3879.
- [13] Sethi, Shilpa, Mamta Kathuria, and Trilok Kaushik. "A Real-Time Integrated Face Mask Detector to Curtail Spread of Coronavirus." *CMES-Computer Modeling in Engineering & Sciences* 127.2 (2021).
- [14] Boulila, Wadii, et al. "A deep learning-based approach for real-time facemask detection." 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2021.
- [15] Sheikh, Burhan ul haque, and Aasim Zafar. "RRFMDS: Rapid real-time face mask detection system for effective COVID-19 monitoring." *SN Computer Science* 4.3 (2023): 288.

- [16] Addagarla, Ssvr Kumar, G. Kalyan Chakravarthi, and P. Anitha. "Real time multi-scale facial mask detection and classification using deep transfer learning techniques." *International Journal* 9.4 (2020): 4402-4408.
- [17] Nasiri, Ehsan, Mariofanna Milanova, and Ardalan Nasiri. "Video Surveillance Framework Based on Real-Time Face Mask Detection and Recognition." *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2021.
- [18] Ai, Mona AS, et al. "Real-time facemask detection for preventing COVID-19 spread using transfer learning based deep neural network." *Electronics* 11.14 (2022): 2250.
- [19] Nasri, Ismail, et al. "MaskNet: Bi-LSTM and LSTM for real-time face mask detection based on deep learning techniques." *International Conference on Digital Technologies and Applications*. Cham: Springer International Publishing, 2021.
- [20] Farkhod, Akhmedov, et al. "Development of Real-Time Landmark-Based Emotion Recognition Bi-LSTM and LSTM for Masked Faces." *Sensors* 22.22 (2022): 8704.
- [21] Kaur, Gagandeep, et al. "Face mask recognition system using Bi-LSTM and LSTM model." *Neuroscience Informatics* 2.3 (2022): 100035.

nafisa final paper

ORIGINALITY REPORT

18%

SIMILARITY INDEX

14%

INTERNET SOURCES

6%

PUBLICATIONS

9%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	journals.itb.ac.id Internet Source	3%
3	Submitted to Huntington Beach Union High School District Student Paper	1%
4	Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, Xu Wu. "Sentiment Analysis of Comment Texts Based on BiLSTM", IEEE Access, 2019 Publication	1%
5	www.researchgate.net Internet Source	1%
6	Submitted to Daffodil International University Student Paper	1%
7	www.wikihow.com Internet Source	<1%
8	fastercapital.com Internet Source	<1%
