MULTI-LABEL HOSTILITY CLASSIFICATION IN BANGLA TEXT

BY

MD RAISUL ISLAM
ID: 201-15-14119

AND

MD ABIR HOSSEN
ID: 201-15-13834

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Shahadat Hossain
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Md Ashraful Islam Talukder
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

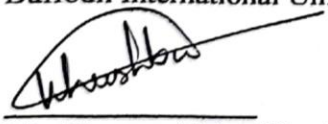
APPROVAL

This Project titled "Multi-Label Hostility Classification in Bangla Text", submitted by **Md Raisul Islam** and **Md Abir Hossen** 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 15-07-2024.

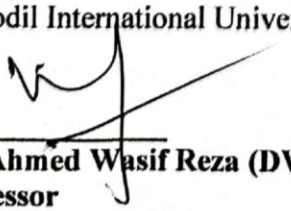
BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah (MIJ) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Assaduzzaman (MA) — **Internal Examiner 1**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Ms. Sharun Akter Khushbu (SAK) — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Ahmed Wasif Reza (DWR) — **External Examiner**
Professor
Department of CSE
East West University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Shahadat Hossain, Lecturer (Senior Scale), Department of Computer Science and Engineering, Daffodil International University.** We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

10.07.24

Shahadat Hossain
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:

10.07.24

Md Ashraful Islam Talukder
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Raisul 10-07-24

Md Raisul Islam
ID: 201-15-14119
Department of CSE
Daffodil International University

Abir 10-07-24

Md Abir Hossen
ID: 201-15-13834
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the final year project successfully.

We are grateful and wish our profound indebtedness to **Shahadat Hossain**, **Lecturer (Senior Scale)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning and Deep Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

There is an immediate demand for automated systems that can identify and handle hostile content due to the quick spread of online communication platforms. This study tackles the problem of multi-label hostility categorization in the context of the Bangla language, where text texts may concurrently include several forms of hostile content. This study offers a thorough method for dealing with the classification of multi-label antagonism in Bangla literature. A deep learning model with various types of hostility is constructed and trained on a large-scale annotated dataset of Bangla text using the latest advances in natural language processing techniques. To capture complex linguistic patterns and contextual connections in Bangla text, the model design combines attention processes with sophisticated neural network components including recurrent and convolutional layers. Additionally, this study looks into a number of pre-processing methods designed especially for the Bangla language, such as addressing typographical variances, word segmentation, and stemming. Furthermore, the study investigates feature engineering techniques to improve the model's capacity to identify subtle language cues that convey animosity in various settings. Experiments carried out on benchmark datasets show how reliable and efficient the suggested method is in correctly identifying various forms of offensive content in Bangla text. We achieved the most optimal result at a training accuracy of 87.67%, and a training loss of 36.89% by using Bi-LSTM model. The model demonstrates its potential utility in practical applications such as sentiment analysis in Bangla-speaking communities, online safety, and content moderation by achieving competitive performance metrics in precision, recall, and F1-score. Ultimately, by providing insights and approaches that may be modified and expanded to handle comparable problems in other languages and cultural situations, this research advances automated hostility detection systems for the Bangla language.

TABLE OF CONTENTS

LISTOF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Overview

Deep The goal of the project is to create a reliable system that can classify hostile information in Bangla text documents. Specifically, the system will handle the problem of multi-label classification, which arises when text documents contain numerous forms of hostile content at the same time. The demand for automated systems that can reliably identify and manage hostile content has increased due to the spread of online platforms and the rise in their prevalence. The work builds a model that can accurately identify different types of hostility in Bangla text by utilizing deep learning techniques and cutting-edge NLP natural language processing techniques. To do this, training and assessment purposes make use of a large-scale annotated dataset of Bangla text that contains a variety of hostile content kinds.

The model architecture is designed to capture intricate linguistic patterns and contextual dependencies present in Bangla text, incorporating components such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention mechanisms. Additionally, the research investigates preprocessing techniques tailored specifically for Bangla text, including word segmentation, stemming, and handling of typographical variations, to ensure optimal performance of the model. Experimental evaluations are conducted on benchmark datasets to assess the performance of the proposed approach in terms of precision, recall, and F1 score. The results demonstrate the efficacy and robustness of the model in accurately identifying multiple types of hostile content in Bangla text across various contexts. Overall, the research contributes to advancing automated hostility detection systems for the Bangla language, offering insights and methodologies that can be applied to enhance online safety and content moderation efforts in Bangla-speaking communities. Moreover, the study lays the groundwork for future research in similar areas, facilitating the development of effective hostility classification systems in other languages and cultural contexts.

1.2 Background and Present State

The rapid expansion of online communication platforms has brought about an alarming rise in hostile content, posing serious threats to online safety and community well-being. In the Bangla-speaking context, there's a pressing need for robust automated systems capable of accurately detecting and managing various forms of hostility, such as hate and cyberbullying. However, existing approaches often struggle to address the complexities of Bangla text and the multi-label nature of hostile content. This research is motivated by the urgency to fill this gap by developing advanced multi-label hostility classification systems tailored specifically for the Bangla language, thereby contributing to safer online environments and fostering inclusive digital communities.

The primary objective of this research is to design, implement, and evaluate an effective multi-label hostility classification system for the Bangla language. This involves developing advanced natural language processing (NLP) techniques and deep learning models tailored to the complexities of Bangla text, with the aim of accurately detecting and categorizing various forms of hostile content. By addressing the multi-label nature of hostility classification and leveraging state-of-the-art methodologies, the research seeks to enhance online safety and content moderation efforts within Bangla-speaking communities.

1.3 Problem Statement

Conducting a study on multi-label hostility classification in Bangla is justified by its societal relevance, linguistic specificity, contribution to NLP research, applications in online safety and moderation, legal compliance, and social and ethical responsibility. By developing effective models for hostility detection in Bangla text, researchers can help create safer, more inclusive, and healthier online environments for Bangla-speaking users. The rationale for conducting a study on multi-label hostility classification in Bangla is driven by several factors:

Addressing Social Concerns: Hostility, aggression, and hate are prevalent issues in online communication platforms and social media, including those in the Bangla language. Analyzing and classifying hostile content in Bangla text can contribute to efforts to mitigate online toxicity, promote civil discourse, and protect individuals from harmful online

interactions.

Protecting User Well-being: Exposure to hostile or aggressive content online can have negative effects on individuals' mental health, well-being, and sense of safety. Developing models to automatically detect and classify hostile content in Bangla text can help identify and filter out harmful content, creating safer and more positive online environments for Bangla-speaking users.

Cultural and Linguistic Specificity: Hostile expressions and hate speech may manifest differently across languages and cultures, making it important to develop models specifically tailored to the linguistic and cultural characteristics of the Bangla language. By focusing on Bangla text, researchers can account for language-specific nuances, slang, cultural references, and linguistic variations that influence the expression and perception of hostility.

Contributing to NLP Research: Multi-label hostility classification in Bangla presents a challenging problem in natural language processing (NLP) and machine learning. Addressing this challenge can advance the state-of-the-art in NLP research, particularly in non-English languages, and contribute to the development of more robust and culturally sensitive AI models for text analysis and classification.

Applications in Online Safety and Moderation: Effective hostility classification models can be integrated into online content moderation systems, social media platforms, and communication tools to automatically detect and flag hostile or abusive content in Bangla text. This can support efforts to enforce community guidelines, prevent harassment and cyberbullying, and maintain a positive online environment for users.

Legal and Regulatory Compliance: Many countries have laws and regulations governing online content, including measures to combat hate speech, discrimination, and incitement to violence. Developing tools for hostility classification in Bangla text can assist online platforms and service providers in meeting legal obligations related to content moderation and user safety.

Social and Ethical Responsibility: Given the potential societal impacts of online hostility and hate speech, researchers have a social and ethical responsibility to develop technologies that promote respectful and inclusive online communication. By addressing the problem of hostility classification in Bangla text, researchers can contribute to efforts to combat online toxicity and promote digital civility and well-being.

1.4 Objectives

The expected output for multi-label hostility classification in Bangla would be a classification model capable of predicting multiple labels representing different types or degrees of hostility present in Bangla language text. The output can vary depending on the specific approach and requirements of the task, but typically includes the following components:

Predicted Labels: The model assigns one or more predefined labels to each input text sample, indicating the presence of specific types of hostility. These labels may include categories hate, harassment, trolling, threats, or offensive language, among others.

Probability Scores: Along with the predicted labels, the model provides probability scores or confidence levels for each predicted label, indicating the model's certainty about the presence of hostility in the text. Higher probability scores suggest greater confidence in the predicted labels.

Multi-label Output Format: The output is presented in a multi-label format, where each predicted label is associated with a binary value indicating its presence or absence in the input text. For example, a binary vector [1, 0, 1, 0, 1] represents the presence of hostility labels at specific positions, with 1 indicating presence and 0 indicating absence.

Thresholding Mechanism: Optionally, the model may incorporate a thresholding mechanism to filter out predicted labels with low probability scores, ensuring that only labels with sufficient confidence are included in the final output. This helps improve the precision of the model's predictions and reduces false positives.

Post-processing Steps: Post-processing steps may be applied to refine the output, such as label smoothing, label filtering, or post hoc adjustments based on domain-specific rules or heuristics. These steps aim to improve the coherence and interpretability of the output labels.

Visualization or Interpretation: Optionally, the output may include visualizations or interpretations of the model's predictions, such as word-level attention maps, saliency scores, or textual explanations highlighting the key features or phrases contributing to each predicted label.

1.5 Scope and Limitations

By addressing these points in this report, a comprehensive overview of the scope and limitations of using CNN, RNN, and BiLSTM models for multi-label hostility classification in Bangla text is provided.

Scope:

Language Focus: Target Language, the primary language of interest is Bangla. This necessitates handling unique linguistic features such as script, syntax, and semantics specific to Bangla. Script and Orthography, focus on text written in Bengali script, which includes handling complex characters and diacritics.

Hostility Categories: Multi-Label Classification, the classification system aims to identify multiple hostility labels such as hate speech, offensive language, abusive content, and other forms of hostile communication within a single instance of text.

Machine Learning Models: CNN, Utilizes Convolutional Neural Networks for feature extraction from text sequences. RNN, Employs Recurrent Neural Networks to capture sequential dependencies in text. BiLSTM, Incorporates Bidirectional Long Short-Term Memory networks to capture context from both past and future states in the text sequence.

Data Sources: Content Types, Includes social media posts, comments, and other user-generated content in Bangla. Data Collection, involves creating or sourcing datasets that are representative of the types of hostility to be classified.

Text Preprocessing: Normalization, standardizes text by handling spelling variations, colloquial language, and diacritics. Tokenization, Splits text into tokens or words while preserving the structure of Bangla. Embedding, uses word embeddings or other representation techniques tailored to Bangla to convert text into numerical form.

Evaluation Metrics: Performance Metrics, employs precision, recall, F1-score, and accuracy to evaluate the effectiveness of the models. Multi-Label Metrics, Uses specific metrics for multi-label classification such as Hamming loss, subset accuracy, and macro scores.

Limitations

Data Availability: Scarcity of Labeled Data, there is a limited availability of labeled datasets for hostility classification in Bangla, which may impact model training and evaluation. Quality of Data, the quality and diversity of the data might be limited, affecting the model's ability to generalize across different contexts.

Class Imbalance: In imbalanced Datasets, Hostile content may be less frequent than non-hostile content, leading to class imbalance issues that can skew model performance. Minority Classes, Difficulty in accurately detecting less frequent hostility categories due to the imbalance.

Language Nuances: Dialects and Informal Usage, Bangla has various dialects and informal usage that can affect the model's understanding and classification accuracy. Contextual Ambiguities, Hostility detection requires understanding the context, which can be challenging for models, especially in cases of sarcasm or indirect hostility.

Resource Intensity: Computational Resources, training complex models like CNN, RNN, and BiLSTM requires substantial computational resources, which might not be accessible to all researchers. Time Consumption, the process of training and fine-tuning these models can be time-consuming.

Overfitting: Model Complexity, Due to the complexity of the architectures and limited data, there is a risk of overfitting where the model performs well on training data but poorly on unseen data. Generalization, ensuring that the model generalizes well to new, unseen data is a significant challenge.

Interpretability: Deep learning models are often seen as black boxes, making it hard to interpret the rationale behind classifications. explanations for why a particular text was classified as hostile can be difficult, impacting trust in the model.

Continuous Learning: The continuous emergence of new hostile expressions and slang necessitates regular updates to the models and retraining with new data. Ensuring that the model adapts to new trends and variations in hostile language over time.

1.6 Report Organization

This layout provides a structured framework for organizing the report content, ensuring clarity, coherence, and accessibility for readers. Our research title is Multi-Label Hostility Classification in Bangla. Our report layout is given

Table of Contents: List of Chapters and Subsections with Page Numbers

Chapter 1: Introduction: Overview of the project objectives, motivation, and significance. Problem statement highlighting the challenges in multi-label hostility classification in Bangla. Scope and objectives of the report.

Chapter 2: Background: Review of existing literature and research related to hostility classification, hostility detection, and natural language processing. Discussion of methodologies, techniques, and models employed in hostility classification for Bengali text and other languages. Identification of gaps, challenges, and opportunities for research in multi-label hostility classification in Bangla.

Chapter 3: Research Methodology: Detailed explanation of the methodology employed for multi-label hostility classification in Bangla. Discussion of feature extraction techniques, including word embeddings, contextual embeddings, and linguistic features. Description of model architectures explored, such as CNNs, RNNs, and transformer-based models. Strategies for handling multi-label classification scenarios and optimizing model performance. Description of the experimental setup, including hardware specifications, software environment, and libraries used. Overview of the training, validation, and testing procedures.

Chapter 4: Experimental Results Presentation of the results of the study, including quantitative and qualitative findings. Description of performance metrics of trained models' accuracy, precision, recall, F1-score, etc. Detailed analysis and interpretation of results, highlighting strengths and limitations of the approach. Comparison of performance with existing methods or baseline models. Discussion of implications of findings and alignment with research objectives. Addressing any challenges encountered during the research.

Chapter 5: Impact on Society: Providing guidelines and recommendations for implementing the hostility classification system in real-world scenarios and content moderation frameworks. Discussion of practical considerations, scalability, and potential

challenges in integrating the system into online platforms. Ethical considerations and implications for societal impact.

Chapter 6: Summary, Conclusion, and Future Directions: Summarizing key findings of the study and their implications. Conclusions drawn from the research and its contributions to the field. Suggestions for future research directions and areas for improvement.

References: Comprehensive list of references cited throughout the report, including academic papers, articles, and relevant literature.

Appendices: Supplementary materials, including Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Analysis, details relevant to the methodology and experimentation. Any additional supporting information deemed necessary for understanding and replicating the research findings.

1.7 Summary

The project aims to develop a system to classify hostile information in Bangla text documents using deep learning and NLP natural language processing techniques. The system will handle multi-label classification, addressing the issue of multiple forms of hostile content in text documents. The model uses recurrent neural networks, convolutional neural networks, and attention mechanisms to capture linguistic patterns and contextual dependencies. Experimental evaluations show the model's efficacy in identifying multiple types of hostile content across various contexts. This research contributes to automated hostility detection systems for Bangla and lays the groundwork for future research in other languages and cultural contexts. The rapid expansion of online communication platforms has led to a rise in hostile content, particularly in the Bangla language. This research aims to develop advanced multi-label hostility classification systems for the Bangla language, addressing the complexities of Bangla text and the multi-label nature of hostile content. The research aims to enhance online safety and content moderation efforts within Bangla-speaking communities, contribute to societal concerns, protect user well-being, and advance natural language processing (NLP) research. It also aims to contribute to legal compliance and social and ethical responsibility by promoting respectful and inclusive online communication. The multi-label hostility classification in Bangla involves a classification model that predicts multiple labels representing different types of hostility in

the text. The output typically includes predicted labels, probability scores, a multi-label format, a thresholding mechanism, post-processing steps, and visualizations. The model uses CNN, RNN, and BiLSTM models to classify hostility, with the primary language being Bangla. The model uses machine learning models like CNN, RNN, and BiLSTM to extract features from text sequences, and uses data sources like social media posts and comments. However, limitations include limited labeled data availability, poor data quality, class imbalance, language nuances, contextual ambiguities, resource intensity, overfitting, interpretability, and continuous learning. These challenges can impact model performance, accuracy, and adaptability to new hostile expressions and slang.

CHAPTER 2

Literature Review

2.1 Preliminaries

Preliminaries for a study on multi-label hostility classification in Bangla include essential steps and considerations before diving into the research process. Clearly articulate the goals and objectives of the study, such as developing effective models for multi-label hostility classification in Bangla, identifying linguistic features indicative of hostility, or assessing the impact of hostility detection on online platforms. Conduct a comprehensive review of existing literature and research related to hostility classification, hostility detection, natural language processing (NLP), and Bengali language processing. Identify relevant methodologies, techniques, models, datasets, and challenges reported in previous studies. Define the problem statement and research questions for the study, highlighting the challenges and nuances specific to multi-label hostility classification in Bangla text. Specify the target labels or categories of hostility to be classified. Identify or collect a suitable dataset of Bangla language text samples annotated with multiple hostility labels. Ensure the dataset is diverse, representative, and sufficiently large to train and evaluate classification models effectively. Develop annotation guidelines and procedures for labeling hostility categories. Consider ethical implications and guidelines for conducting research on sensitive topics like hostility classification. Address issues such as privacy protection, bias mitigation, fairness, and responsible AI practices throughout the research process. Define the methodology and approach for multi-label hostility classification in Bangla. Determine feature extraction techniques, model architectures, training strategies, and evaluation metrics to be employed. Consider the suitability of existing NLP tools and libraries for processing Bangla text. Set up the experimental environment, including hardware specifications, software tools, programming languages, and libraries required for data preprocessing, model development, training, and evaluation. Ensure compatibility with Bangla language processing tools and resources. Establish baseline models or methods for comparison to evaluate the performance of proposed approaches. This may include simple rule-based systems, traditional machine learning classifiers, or existing state-of-the-art models for hostility classification in other languages. Identify potential collaborators, advisors, or experts in the field of Bangla language processing, NLP, and hostility detection. Seek access to relevant resources,

datasets, tools, or computational infrastructure necessary for conducting the study. Develop a timeline or project plan outlining key milestones, tasks, and deadlines for each phase of the research, from data collection and model development to experimentation, analysis, and reporting.

2.2 Related Works

Sarker, M., Hossain et al. [1] The rise of social media in Bangladesh has led to increased hate speech, cyberbullying, and online harassment. To address these issues, research is being conducted to create a dataset of Bengali comments from social platforms and develop a model to quickly detect anti-social comments. Utilizing 2000 comments from Facebook and YouTube, various machine learning techniques including artificial neural networks and supervised classifiers were employed. This research aims to prevent anti-social activities in the Bengali community, filling a gap in existing studies on anti-social classification in the Bangla language. Bhardwaj et al. [2] The passage highlights the growing concern over the proliferation of hostile content online, particularly on social media platforms, exacerbated by recent events like the COVID-19 pandemic, Black Lives Matter, and #MeToo movements. While existing systems address aspects of hostile post detection, there's a lack of unified approaches for addressing multiple dimensions of hostility. Moreover, research predominantly focuses on the English language, neglecting regional languages like Hindi and Bengali. The paper introduces HostileNet, a deep learning framework utilizing HindiBERT contextual representations and hand-crafted lexicon features for hostile post classification in Hindi. It also proposes a novel mechanism for fine-tuning attention vectors for each hostile dimension. Evaluation on CONSTRAINT-2021's multi-label Hindi dataset demonstrates HostileNet's superior performance compared to existing systems, including the best-performing system in the CONSTRAINT-2021 shared task. The paper includes thorough analyses of the results, including ablation studies, error analysis, attention heatmap analysis, and lexicon feature analysis. De, A., Elangovan et al. [3] The passage discusses the need for hostility detection on social media platforms, particularly in resource-constrained languages like Hindi. It presents a neural network-based approach utilizing pre-trained multilingual Bidirectional Encoder Representations of Transformer (mBERT) for hostility detection in Hindi posts. Extensive experiments were conducted, exploring various pre-processing techniques, pre-trained models, and neural

architectures. The best-performing model achieved high F1 scores for hostile, fake, hate, offensive, and defamation labels, outperforming existing baseline models and establishing itself as state-of-the-art for hostility detection in Hindi posts. Shakil, M. H. et al. [4] The passage highlights the rapid evolution of data innovation and the increasing popularity of social media platforms like Facebook, Twitter, and Instagram, particularly in Bengali-speaking regions like Bangladesh. Unfortunately, these platforms have become hubs for toxic remarks and cyberbullying. Despite numerous studies, accurate results have been elusive. The study proposes using pre-trained transformer models for early identification of malicious Bengali text using simple Natural Language Processing (NLP) techniques. It suggests a Convolutional Neural Network with a Bi-Directional Long Short-Term Memory (CNN-BiLSTM) hybrid strategy for classification, capable of categorizing Bengali text into six levels. Furthermore, conventional Machine Learning methods and Explainable AI techniques are employed, with a Stacking Classifier ensemble approach being superior in performance. Nafis et al. [5] This paper addresses the serious issue of cyberbullying, which is prevalent on social networking sites. The lack of high-quality training data, especially for multilingual and low-resource scenarios, presents a challenge in accurately detecting cyberbullying. The paper introduces a manually annotated dataset of 10,000 English and Hindi-English code-mixed tweets for aggression and offensive language detection. The annotations demonstrate substantial agreement among annotators. Pre-trained language models are fine-tuned using this dataset, achieving macro F1-scores of 67.87 and 65.45 on aggression and offensive language detection tasks. Cross-dataset transfer learning is conducted to benchmark the dataset against existing ones. The paper also provides a detailed analysis of prediction errors and publicly releases the dataset, code, and models. Hossain et al. [6] The paper discusses the exponential growth of digital Bengali text content online and the challenges it poses in terms of organization and manipulation. It proposes an intelligent text document classification system to address this issue. This task is particularly challenging due to the lack of linguistic resources, NLP tools, and large text corpora in Bengali. The proposed model utilizes GloVe embedding and a Very Deep Convolution Neural Network (VDCNN) classifier. To overcome the scarcity of standard corpora, the paper develops two datasets: Embedding Corpus (EC) and Bengali Text Classification Corpus (BDTC). Additionally, it introduces the Embedding Parameters Identification (EPI) Algorithm to select optimal embedding parameters for low-resource

languages like Bengali. Evaluation of various embedding models shows GloVe to be the most suitable for Bengali text. Experimental results demonstrate that the proposed GloVe + VDCNN model outperforms other classification models, achieving the highest accuracy of 96.96% in Bengali text classification. Naha et al. [7] The passage discusses the popularity of micro-blogging platforms like Tumblr, Twitter, and others, where users share their thoughts and engage in discussions daily. The research aims to conduct sentiment analysis to extract emotions from textual content using various tracking methods. Multiple classifiers are trained to identify basic moods such as Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation from text. The research seeks to provide an overview of sentiment analysis techniques and related areas, offering insights into understanding emotions expressed in textual content on micro-blogging platforms. Sazzed et al. [8] The paper addresses the escalating concern over abusive and vulgar language in social media, particularly in low-resource languages like Bengali, which has received limited research attention. It conducts the first comprehensive analysis of vulgarity in Bengali social media content, creating benchmark corpora of 7,245 reviews from YouTube, manually annotated for vulgar and non-vulgar categories. Findings indicate a significant presence of vulgar language, ranging from 20% to 34% across the corpora. Various approaches, including classical machine learning, deep learning, and lexicon-based methods, are employed for the automatic identification of vulgarity. The study observes that lexicon-based methods are effective due to the prevalence of swear terms. The deep learning-based BiLSTM model demonstrates the highest recall scores for identifying vulgarity. Additionally, the analysis highlights the strong correlation between vulgarity and negative sentiment in social media comments. Sazzed et al. [9] The passage discusses the absence of tools and resources for detecting profane and obscene language in Bengali, a low-resource language. It introduces a Bengali obscene lexicon comprising over 200 terms representing filthy, slang, profane, or obscene language. The lexicon is developed using a semi-automatic methodology, utilizing an obscene corpus, word embedding, and part-of-speech taggers. The evaluation shows the lexicon achieves around 85% coverage for detecting obscene and profane content in Bengali social media. Experimental results suggest the lexicon's effectiveness in identifying obscenity in Bengali social media content. Sinha et al. [10] study analyzes English-Hindi (Hinglish) content on social media platforms like Facebook, focusing on multi-label emotion classification, which considers the coexistence of multiple emotions

within a single instance. 15,995 code-mixed Facebook status updates are annotated with nine emotions. Various pre-processing techniques are applied to improve accuracy, and five multi-level classification algorithms are tested. Results show that status updates can evoke multiple emotions, with the Classifier Chains algorithm achieving the highest accuracy of 86% due to its ability to consider correlations between emotions. The study's findings have implications for decision-making processes based on emotional analysis of text.

2.3 Comparison between existing works

The comparative analysis of the related work for multi-level hostility classification in the Bangla language reveals several areas requiring further exploration. Firstly, existing studies primarily focus on hostility classification in languages like English and Hindi, with limited attention given to Bangla. Secondly, while some techniques have been adapted to Bangla, there is a lack of comprehensive approaches tailored specifically to its linguistic nuances and cultural context. Additionally, there is a scarcity of annotated datasets for Bangla, hindering the development and evaluation of effective models. Furthermore, there's a need for research addressing the multi-level nature of hostility, as existing studies often focus on binary classification.

TABLE 2.3: COMPARATIVE ANALYSIS WITH PREVIOUS WORK

SLNo	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Sarker, M., Hossain et al. [1]	Logistic Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM)	RF=71.28% and SVM70.26%. LR=74.36% and MNB=80.51%. ANN=78.89%
2.	Bhardwaj et al. [2]	HindiBERT	Highest 97.51% F1
3.	De, A., Elangovan et al. [3]	mBERT	obtained 92.60%, 81.14%,69.59%,

			75.29% and 73.01%
4.	Shakil, M. H. et al. [4]	CNN-BiLSTM	_____
5.	Nafis et al. [5]	PTLMs	F1-scores of 67.87 and 65.45
6.	Hossain et al. [6]	GloVe + VDCNN	Highest 96.96% accuracy
7.	Naha et al. [7]	KNN, SVM, Naïve Bayes and Decision Tree	accuracy of 90%
8.	Sazzed et al. [8]	BiLSTM, CML, SGD	recall score of 0.85
9.	Sazzed et al. [9]	Lexicon, LR, SVM,SGD	_____
10.	Sinha et al. [10]	ML-Knn, Lexicon	the highest accuracy of 86%

2.4 Open Issues

The scope of the problem for multi-label hostility classification in Bangla encompasses various dimensions related to identifying and addressing hostile or abusive content in Bangla language text. Here's our outline of the scope:

Language Specificity: The problem focuses specifically on hostility classification in Bangla language text, considering the linguistic characteristics, nuances, and cultural context unique to the Bengali language. This includes variations in dialects, regional expressions, and socio-cultural factors that influence the manifestation of hostility in Bangla text.

Multi-label Classification: Unlike binary classification tasks that categorize text as either hostile or non-hostile, multi-label hostility classification involves assigning multiple labels or categories to each text sample, representing different types or degrees of hostility. The scope encompasses identifying various forms of hostility, such as hate, harassment, insults,

threats, or offensive language, among others.

Textual Domains: The problem extends across different domains of Bangla language text, including social media, online forums, news articles, blogs, comment sections, and other digital communication platforms. Hostile content may manifest differently in each domain, requiring adaptive classification models capable of addressing diverse contexts.

Data Collection and Annotation: Collect data from social media comments and annotated with multiple hostility labels. Islam, Khondoker Ittehadul, et al. [22] Employed ten undergraduate students and provided them with detailed annotation guidelines and use majority voting to assign the final class label. We follow this annotation procedure to annotate our dataset. This includes developing annotation guidelines, procedures, and quality control measures to ensure consistency and reliability in labeling hostile content.

Model Development: The problem encompasses developing effective machine learning or deep learning models for multi-label hostility classification in Bangla. This includes exploring various model architectures, feature extraction techniques, and training strategies tailored to the linguistic and cultural characteristics of Bangla text.

Performance Evaluation: The scope involves evaluating the performance of hostility classification models using appropriate metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Evaluation should consider the effectiveness of models across different hostility categories and domains of Bangla text.

Practical Applications: The scope extends to practical applications of multi-label hostility classification in Bangla, including content moderation on online platforms, social media monitoring, hate detection, user safety measures, and law enforcement efforts to combat online harassment and abuse.

TABLE 2.4: ISSUES AND CHALLENGES WITH STRATEGIES PLANS

S.No.	Issues and Challenges	Strategies or Plans
1	Language and Cultural Context	We will develop and use datasets that are representative of various Bangla dialects and cultural contexts to train models that can understand regional variations and subtleties.

2	Annotation and Labeling Inconsistencies	Using multiple annotators for each piece of content minimizes personal bias and improves the reliability of the labels through consensus or majority rules.
3	Data Scarcity and Imbalance	We will also utilize techniques like synonym replacement, back-translation, and other text-generation methods to increase the volume of training data for underrepresented classes.
4	Adaptation and Evolution	We will utilize active learning techniques that prioritize data points for maximum learning and may involve human annotators.

2.5 Summary

A study on multi-label hostility classification in Bangla involves defining goals, conducting a literature review, defining problem statements, collecting a suitable dataset, and considering ethical implications. The methodology and approach for the study include feature extraction techniques, model architectures, training strategies, and evaluation metrics. The study also addresses the growing concern over the proliferation of hostile content online, particularly in regional languages like Hindi and Bengali. Previous research has introduced HostileNet, a deep learning framework for hostile post classification in Hindi, and De, A., Elangovan et al., presents a neural network-based approach for hostility detection in Hindi posts. The study focuses on multi-level hostility classification in the Bangla language, highlighting the need for further exploration. Existing studies mainly focus on English and Hindi, with limited attention given to Bangla. There is a lack of comprehensive approaches tailored specifically to its linguistic nuances and cultural context. There is also a scarcity of annotated datasets for Bangla, hindering the development and evaluation of effective models. The problem for multi-label hostility classification in Bangla encompasses various dimensions related to identifying and addressing hostile or abusive content in Bangla language text. The study aims to develop effective machine learning or deep learning models, evaluate their performance, and apply them in practical applications such as content moderation, social media monitoring, hate detection, and law enforcement efforts.

CHAPTER 3

Methodology

3.1 Overview

The research subject for multi-label hostility classification in Bangla focuses on developing and evaluating machine learning or deep learning models capable of identifying and categorizing various forms of hostile or abusive content in Bangla language text. The primary subject of the research is the classification of Bangla language text into multiple categories or labels representing different types or degrees of hostility. This includes identifying and categorizing various forms of hostile content, such as hate, harassment, threats, trolling, or offensive language, among others. Collection or curation of a diverse and representative dataset of Bangla language text samples annotated with multiple hostility labels. Development of annotation guidelines and procedures for labeling hostile content, ensuring consistency and reliability in annotation.

Extraction of word embeddings or contextual embeddings from Bangla text using pre-trained language models or custom embedding techniques. Identification and extraction of linguistic features relevant to hostility classification, such as sentiment analysis, syntactic patterns, semantic cues, and cultural context.

Model Development: Exploration and development of Deep Learning Architectures. Experimentation with deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bi-LSTMs models for multi-label hostility classification. Evaluation of model performance using metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Analysis of classification results using confusion matrices to understand model behavior across different hostility categories.

3.2 Proposed Methodology

The research work on multi-level hostility classification in the Bangla language employs a systematic methodology to effectively address the complexities inherent in analyzing hostile content across multiple levels. The methodology encompasses several key stages, each designed to facilitate the accurate classification of hostile language while considering linguistic nuances specific to the Bangla language context.

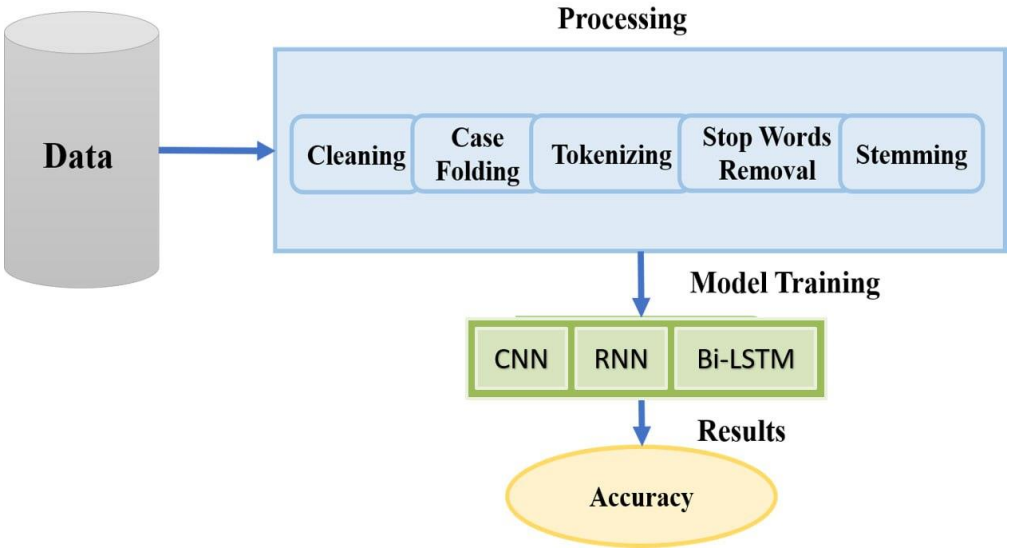


Figure 3.2.1: Workflow diagram for Bangla Hostility analysis

We will prepare our Bangla text dataset, ensuring it is annotated with multiple hostility labels for each text sample. Split the dataset into training, validation, and testing sets to facilitate model training, tuning, and evaluation. Then tokenize Bangla text into individual words or characters. Convert tokens into numerical representations, possibly using word embeddings like Word2Vec, FastText, or pre-trained language models. Design a BiLSTM architecture capable of processing sequential data bidirectionally. Stack multiple BiLSTM layers to capture increasingly abstract representations of the input text. Optionally, incorporate additional layers such as convolutional layers or attention mechanisms to enhance model performance and interpretability. Configure the output layer to produce multi-label predictions, typically using a sigmoid activation function to enable independent classification of each hostility category. Evaluate the trained BiLSTM model on the held-out test set using evaluation metrics suitable for multi-label classification tasks. Common metrics include F1 Score, validation loss, Precision, Recall, and Mean Average Precision (mAP). Interpret model performance across different hostility categories and analyze error patterns to identify areas for improvement. By incorporating a BiLSTM model into our methodology, we will leverage its ability to capture long-range dependencies and contextual information in Bangla text, thereby enhancing the accuracy and effectiveness of multi-label hostility classification tasks.

The Bi-LSTM model assigns equal importance to all inputs. The hostility polarity of the text in hostility analysis is mostly determined by the presence of words that convey hostility information. This paper implements hostility reinforcement of hostility word vectors.

hostility analysis tasks are considered 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 hostility 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 hostility inclination of the comments is acquired. The activation function utilized in feed forward 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.

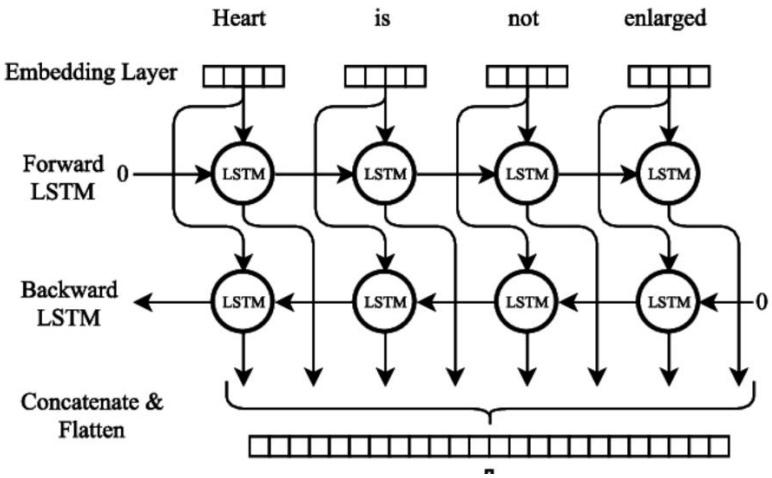


Figure 3.2.2: Bi-LSTM architecture for hostility

The schematic diagram illustrating the hostility technique described in this paper can be found in Figure 3.2.2 The left subgraph represents the process of extracting features from comment text. The right subgraph refers to the process of determining the hostility polarity of the comment text. NodeNum represents the number of nodes in the hidden layer of the LSTM.

3.3 Hardware and Software Requirement

For Setting up experiments for multi-label hostility classification in Bangla involves configuring the environment, preparing the dataset, defining the methodology, and conducting the experiments. Here's a detailed outline of the experimental setup:

Hardware Requirements: Determine the hardware infrastructure needed for model

training and evaluation. This may include CPUs, GPUs, or TPUs with sufficient memory and computational power.

Software Environment: Choose the programming language and deep learning framework for model development. Common choices include Python with TensorFlow, PyTorch, or Keras. Install necessary libraries and packages for data preprocessing, model development, and evaluation of scikit-learn, pandas, and NumPy.

Dataset Preparation: Collect or curate a dataset of Bangla language text samples annotated with multiple hostility labels. Ensure the dataset is diverse, balanced, and representative of different hostility categories. Preprocess the dataset by tokenizing text, removing stopwords, performing stemming or lemmatization, and encoding labels into a suitable format for model training.

Feature Extraction: Choose appropriate feature extraction techniques to represent Bangla text data in a format suitable for model training. This may include words, contextual embeddings, or linguistic features. Precompute word embeddings or extract features from text data using pre-trained language models or custom pipelines.

Model Development: Experiment with different model architectures suitable for multi-label text classification, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bidirectional LSTMs (Bi-LSTMs). Define the architecture, hyperparameters, and optimization algorithms for each model variant to be evaluated.

Training Procedure: Train the hostility classification models on the training dataset using the defined architecture and hyperparameters. Monitor training progress, track performance metrics on the validation set, and implement early stopping criteria to prevent overfitting. Save checkpoints of trained models at regular intervals for later evaluation and analysis.

Evaluation: Evaluate the trained models on the test dataset using selected evaluation metrics accuracy, precision, recall, and F1-score. Perform cross-validation to assess the generalization and robustness of models across different folds. Analyze model performance, identify strengths and weaknesses, and compare results across different model variants.

3.4 Project Management and Financial Analysis

Project Management: Form a multidisciplinary team comprising researchers, linguists,

data annotators, and machine learning engineers proficient in Bangla language and natural language processing (NLP). Develop a detailed project timeline with milestones for data collection, preprocessing, model development, evaluation, and deployment. Assign specific tasks to team members based on their expertise and create a streamlined workflow to ensure efficient progress. Establish regular meetings and communication channels to facilitate collaboration and address any challenges promptly

Project Management Plan: We develop this project plan for Multi-label Hostility Classification in the Bangla Language. Our project management plan will follow these steps.

Project Objectives:

- Develop a multi-label hostility classification system for Bangla text.
- Enable accurate identification and classification of hostile content expressed in the Bangla language.
- Provide a valuable tool for content moderation, sentiment analysis, and online safety in Bangla-language platforms and communities.

Project Timeline:

Phase 1: Project Planning and Research (2-4 weeks)

Phase 2: Data Collection and Annotation (4-6 weeks)

Phase 3: Model Development and Training (6-8 weeks)

Phase 4: Evaluation and Fine-tuning (4-6 weeks)

Phase 5: Deployment and Integration (2-4 weeks)

Phase 6: Testing and Quality Assurance (4-6 weeks)

Phase 7: Post-Launch Support and Maintenance (Ongoing)

Resource Planning:

Equipment and Tools

- High-performance computing resources for model training.
- Software tools for data preprocessing, machine learning, and evaluation.
- Collaboration platforms for team communication and coordination.

Software and Technologies

- Natural Language Processing (NLP) libraries and frameworks suitable for Bangla text.
- Machine learning algorithms for text classification.
- Cloud platforms for scalable computing and storage.

Data and Content

- Bangla text datasets with annotated hostility labels.
- Annotation tools for manual labeling and validation.

Training and Skill Development

- Ensure team members possess expertise in Bangla language, NLP, machine learning, and software development.
- Provide training on specific tools, techniques, and methodologies required for the project.

Risk Management:

Conduct a thorough risk assessment, considering potential challenges related to data availability, annotation quality, model bias, and regulatory compliance. Implement risk mitigation strategies, such as diversifying data sources, conducting sensitivity analyses, and incorporating fairness and transparency measures into the model. Monitor project progress closely and address any emerging risks promptly to safeguard project objectives and stakeholder interests. Here is a SWOT analysis for the e-commerce for Multi-label Hostility Classification in the Bangla Language.

Figure 3.4.1: SWOT analysis for Multi-label Hostility Classification in the Bangla Language

Communication Plan:

- Regular Team Meetings Weekly or bi-weekly meetings to discuss progress, address challenges, and align on priorities.
- Stakeholder Updates Provide regular updates to project stakeholders on milestones achieved, challenges encountered, and future plans.
- Documentation and Reporting Maintain detailed documentation of project activities, including data sources, methodology, experimental results, and decisions made. Generate progress reports and summaries for stakeholders as needed.
- Feedback Channels Establish channels for collecting feedback from team members, stakeholders, and subject matter experts throughout the project lifecycle to inform decision-making and improve project outcomes.

Financial Analysis:

TABLE 3.4: FINANCIAL ANALYSIS TABLE FOR OUR RESEARCH PROJECT

SN	**Components**	**Estimated Cost (BDT)**
01.	Infrastructure	2000
02.	Visiting Stakeholders	1000
03	Software and Tools	1500
04.	Data Collection and Processing	2000
05.	Documentation and Report Writing	500
06.	Software licenses, tools licenses	1500
09.	Data scientists	8,00
10.	Annotators	400
13.	Model Training	700
14.	Contingency (10% of total)	1000
	Total Estimated Cost	11,400

This table provides an overview of the estimated costs for each expense category involved in the research project. Actual costs may vary depending on factors such as project duration, specific requirements, and market rates. Adjustments can be made based on available resources and funding opportunities.

3.5 Summary

The research focuses on developing and evaluating machine learning models for multi-label hostility classification in Bangla language. The primary goal is to identify and categorize hostile content in Bangla language text into different labels, such as hate, harassment, threats, trolling, or offensive language. The research involves collecting a diverse dataset of Bangla language text samples annotated with multiple hostility labels,

developing annotation guidelines, extracting word embeddings, and identifying linguistic features relevant to hostility classification. Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bi-LSTMs are used for model development. The methodology includes several key stages to facilitate accurate classification of hostile language while considering linguistic nuances specific to the Bangla language context. The proposed methodology leverages the ability of BiLSTM models to capture long-range dependencies and contextual information in Bangla text, enhancing the accuracy and effectiveness of multi-label hostility classification tasks. The research focuses on developing a multi-label hostility classification system for Bangla text, enabling accurate identification and classification of hostile content expressed in the Bangla language. The project involves configuring the environment, preparing a dataset, defining the methodology, and conducting experiments. The hardware requirements include CPUs, GPUs, or TPUs, and the software environment includes Python with TensorFlow, PyTorch, or Keras. The dataset is annotated with multiple hostility labels, and feature extraction techniques are chosen. The model development process involves experimenting with different model architectures, training the models, and evaluating them. The project management involves forming a multidisciplinary team, developing a project timeline, and establishing regular meetings and communication channels. The project's financial analysis shows an estimated cost of 11,400, which includes infrastructure, visiting stakeholders, software and tools, data collection and processing, documentation and reporting, and contingency.

CHAPTER 4

Implementation

4.1 Overview

For this research work, we need methodology requirement analysis. In the research on multi-level hostility classification in the Bangla language, a comprehensive requirement analysis serves as the foundation for designing an effective methodology tailored to the unique linguistic and cultural characteristics of the Bangla language. The requirement analysis encompasses several key aspects essential for the successful development and implementation of a robust hostility classification system:

Complexity Understanding: It is imperative to comprehend the linguistic complexity inherent in Bangla language expressions, including dialectical variations, idiomatic expressions, and cultural nuances. This understanding forms the basis for devising appropriate feature engineering techniques and model architectures capable of capturing the subtleties of hostile language in Bangla.

Data Availability and Diversity: Access to a diverse and representative dataset of Bangla text samples spanning multiple levels of hostility is essential. The dataset should encompass a wide range of sources, including social media, online forums, news articles, and user-generated content, to ensure comprehensive coverage of hostile language variations encountered in real-world scenarios.

Annotation and Labeling Guidelines: Clear and consistent annotation guidelines are necessary for annotating the dataset with appropriate hostility labels corresponding to different levels of aggression, hostility, or toxicity. These guidelines should account for cultural sensitivities and contextual factors specific to the Bangla language, ensuring accurate and reliable labeling of the dataset for training and evaluation purposes.

Feature Selection: Identification of relevant linguistic features and extraction criteria tailored to the characteristics of the Bangla language is crucial. This involves selecting linguistic cues, syntactic structures, semantic patterns, and contextual markers indicative of hostile intent or sentiment in Bangla text. The feature selection process should prioritize discriminative features capable of distinguishing between different levels of hostility effectively.

Model Selection and Evaluation Metrics: The selection of appropriate machine learning

and deep learning models for hostility classification in the Bangla language requires careful consideration. Models should be chosen based on their suitability for handling text data, scalability, and performance across multiple hostility levels. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC should be selected to assess the classification performance comprehensively.

Ethical and Sociocultural Implications: Ethical considerations about content moderation, privacy, and freedom of expression must be carefully addressed throughout the methodology design and implementation process. Cultural sensitivities, biases, and potential implications of automated content moderation in Bangla language domains should be thoroughly evaluated to mitigate unintended consequences and ensure responsible use of the classification system.

4.2 Train Model

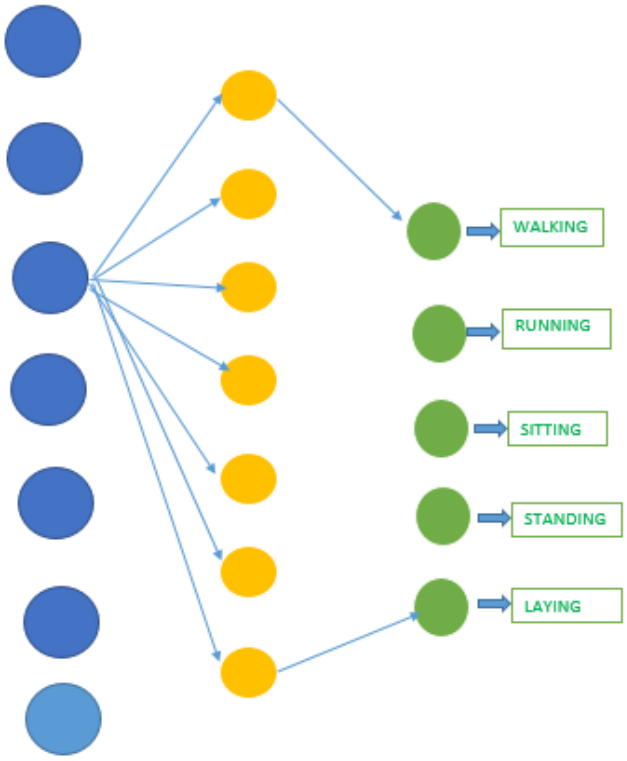


Figure 4.2.1: CNN hostility classifier

Figure 3.4.3 shows the CNN hostility 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 hostility classification system is shown in Figure 3.4.3 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.

$$S(x) = \frac{1}{1+\mathrm{ex}} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

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 3.4.4 illustrates the application of Recurrent Neural Network (RNN) in Chinese text hostility 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 (w1, w2, w3, w4), which is subsequently fed into the RNN in a sequential manner as the matching word

vector (wt, wt, wt, wt).

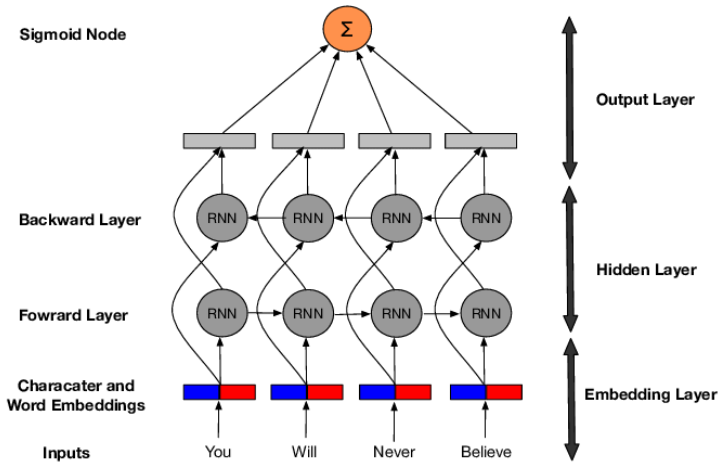


Figure 4.2.2: The RNN architecture for the hostility analysis model.

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 3.4.5.

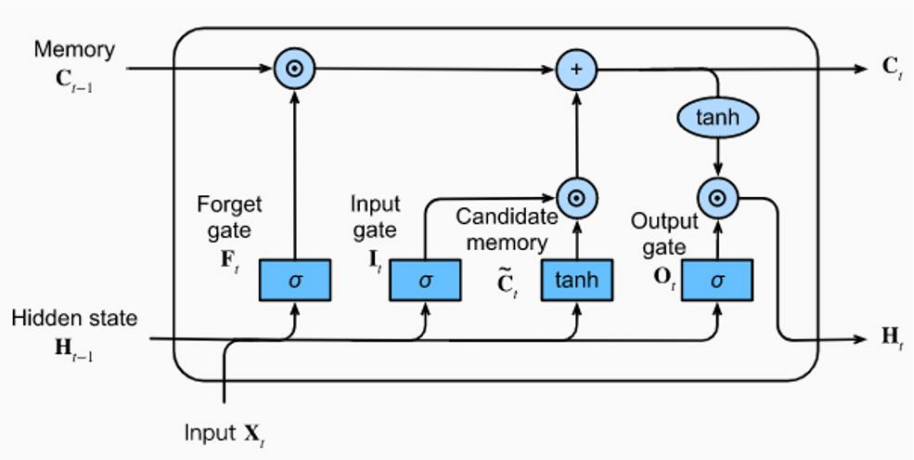


Figure 4.2.3: Diagram of LSTM structure for hostility.

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

In both the traditional recurrent neural network model and the 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.

4.3 System Testing

Hostility 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 hostility. 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 hostility detection. Address the issue of model interpretability in the context of sensitive hostility. 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.

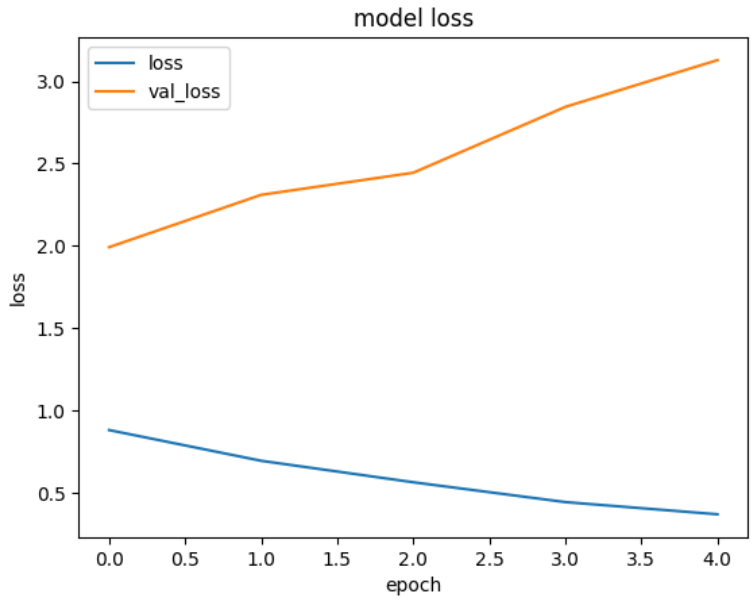


Figure 4.3.1: Loss and val_loss comparison graph

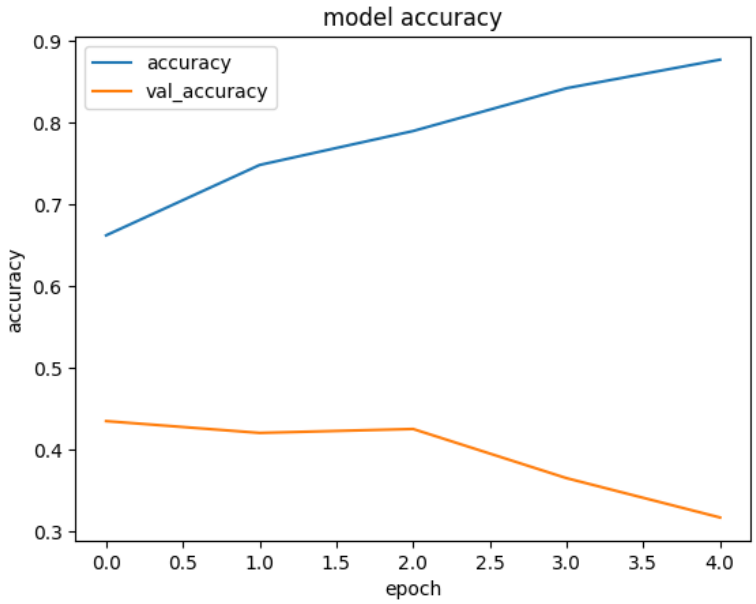


Figure 4.3.2: Accuracy and val_accuracy comparison graph

Figures 4.3.1 and 4.3.2show a comparison line between the accuracy, loss, and validation accuracy, validation loss. Consider the scalability and real-time deployment of the CNN Base model for hostility 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 hostility 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 hostility detection and classification using the CNN Base model. Highlight the potential impact on content moderation, hostility 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 hostility 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.

4.4 Summary

The CNN hostility classifier architecture is a multilayer perceptron model that uses convolution processes to produce accurate results. It was first created for text image recognition and has been demonstrated to produce very accurate results. The one-dimensional convolutional layer of the Bi-LSTM and LSTM architecture uses sentences as input, with three filters, each with a batch size of 128. The dimensionality of the convolution output is decreased by a max pooling layer of size 3, followed by the flattened layer, which is then followed by a fully linked layer made up of softmax activation function layers. 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, consisting of nodes connected between hidden layers and having the ability to dynamically learn sequence features. This research proposes a hostility analysis approach for comments based on BiLSTM to address the limitations of current comment hostility analysis methods.

The Bi-LSTM model combines bidirectional recurrent neural network (BiRNN) models with LSTM units to effectively capture context information. The study discusses the key findings, implications, limitations, and future directions of the research on hostility classification using Convolutional Neural Networks and Bi-LSTM. It also discusses the performance of the CNN, Bi-LSTM, and LSTM models in detecting and classifying hostility, their generalization performance, and potential challenges in adapting the model to new or unseen data.

CHAPTER 5
Result and Analysis

5.1 Overview

The study focuses on setting up experiments for multi-label hostility classification in Bangla, which involves configuring the environment, preparing the dataset, defining the methodology, and conducting the experiments. The hardware requirements include CPUs, GPUs, or TPUs with sufficient memory and computational power. The software environment includes Python with TensorFlow, PyTorch, or Keras, as well as libraries and packages for data preprocessing, model development, and evaluation. Dataset preparation involves collecting or curating a dataset of Bangla language text samples annotated with multiple hostility labels. The dataset is diverse, balanced, and representative of different hostility categories. Feature extraction techniques are chosen to represent Bangla text data in a suitable format for model training. Model development involves experimenting with different model architectures suitable for multi-label text classification, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bidirectional LSTMs (Bi-LSTMs). The training procedure involves training the hostility classification models on the training dataset using the defined architecture and hyperparameters.

Evaluation is performed on the test dataset using selected evaluation metrics, including accuracy, precision, recall, and F1-score. Cross-validation is performed to assess the generalization and robustness of models across different folds. The experiment used a dataset of 2080 Facebook Bangla hostility comments from real-time Facebook comments, built with the Keras library for Python. The dataset was split into 80% for the training set and 20% for the validation set, and the results are shown in Figure 5.2.2.

5.2 Experimental

While in this paper, we use a dataset of a total of 2080 Facebook Bangla hostility comments and collected each data from real-time Facebook comments. The method for hostility classification in the Indonesian language using Bi-Lstm was built with the Keras library for Python. The dataset has 6 hostility class labels ('trolling', 'non-hostile', 'offensive',

'misinformation', 'hate', and 'harassment').

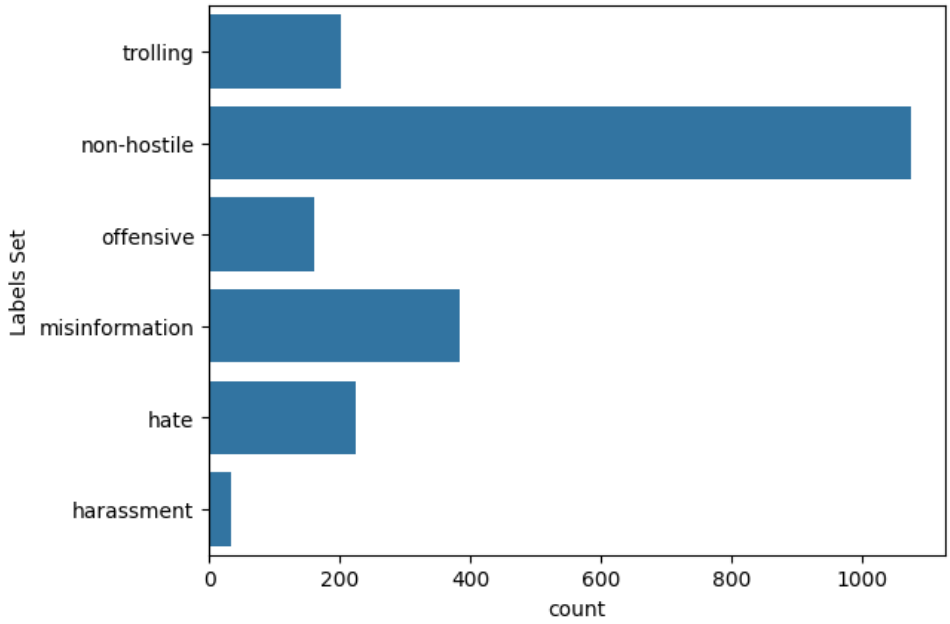


Figure 5.2.1: Hostility class labels of our dataset

Preprocessing Steps:

Here we discuss our data preprocessing step by step:

text_to_word_list(text) function: In this function, we take a string text as input and split it into a list of words. It uses the split () method to split the text based on whitespace and returns the resulting list of words.

replace_strings(text) function: In this function, we replace certain patterns in the input text with an empty string. Specifically, it removes emojis and non-Bangla characters. It uses regular expressions (re module) to define patterns for emojis and English characters and then applies the sub () method to replace these patterns with an empty string.

remove_punctuations(my_str) function: By this function, we remove punctuations from the input string my_str. It defines a string punctuation containing all punctuation characters, then iterates over each character in my_str and appends it to no_punct if it's not in the punctuation string.

joining(text) function: With this function, we take a list of words `text` and join them back into a single string using whitespace as the separator.

preprocessing(text) function: In this function, we apply preprocessing steps to the input text text. It first removes punctuations and then replaces certain strings (emojis and non-Bangla characters) using the replace_strings () function.

5.3 Performance

In this section, I discuss the performance, implications, and future directions of the study. Discuss the performance of the CNN, Bi-LSTM, and LSTM models in detecting and classifying hostility. 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 hostility detection.Address the issue of model interpretability in the context of sensitive hostility. 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.The dataset used for training was split into 80% (1664 comments) for the training set and 20% (416 comments) for the validation set. The experiment was conducted for 1, 2, 3, 4,and 5 epochs. The experimental results are shown in Figure5.2.2.

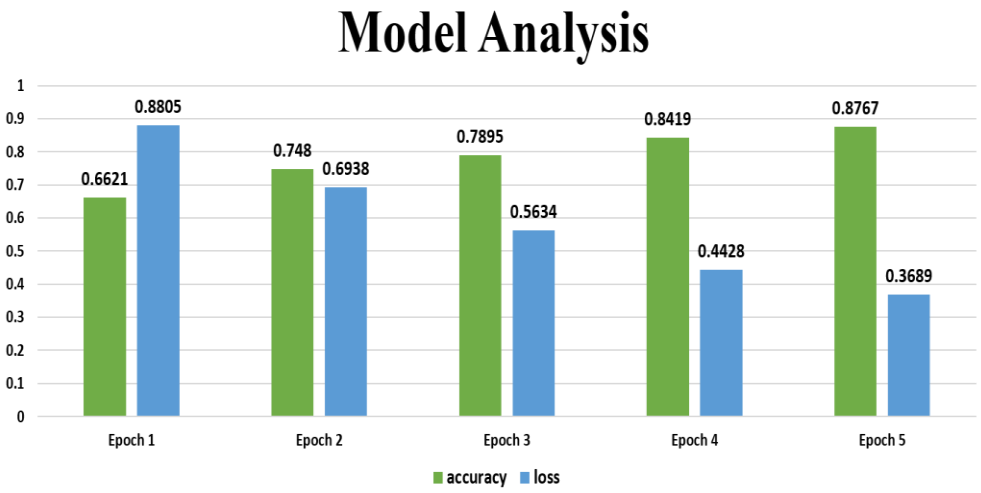


Figure 5.3.1: Model Analysis results of our dataset

Figure 5.2.2 shows the 5th epoch had the highest accuracy result. However, the loss started to increase gradually after 2 epochs, while the training loss kept decreasing, which caused

the model to overfit when trained longer. Both training loss by the 3rd epoch hit the lowest percentage. The model checkpoint stopped saving at 5 epochs, therefore it achieved the most optimal result at a training accuracy of 87.67%, a training loss of 36.89%.

```
[36] class_names = ['trolling', 'non-hostile', 'offensive', 'misinformation', 'hate', 'hara

from sklearn.metrics import classification_report

predictions = model.predict(testing_padded)

predicted_labels = (predictions > 0.5).astype(int)

print(classification_report(test_labels, predicted_labels, target_names=class_names))
```

```
13/13 [==============================] - 5s 376ms/step
                precision    recall  f1-score   support

      trolling       0.05      0.17      0.08        40
   non-hostile       0.53      0.63      0.58       221
     offensive       0.08      0.52      0.15        33
misinformation       0.18      0.73      0.29        75
          hate       0.11      0.31      0.16        36
    harassment       0.03      0.09      0.05        11

     micro avg       0.22      0.55      0.32       416
     macro avg       0.17      0.41      0.22       416
  weighted avg       0.34      0.55      0.39       416
   samples avg       0.26      0.55      0.34       416
```

Figure 5.3.2: Classification report of our dataset

Figure 5.2.3 shows the classification report of our dataset. Here we show the precision, recall, and f1-score of our dataset according to our class labels.

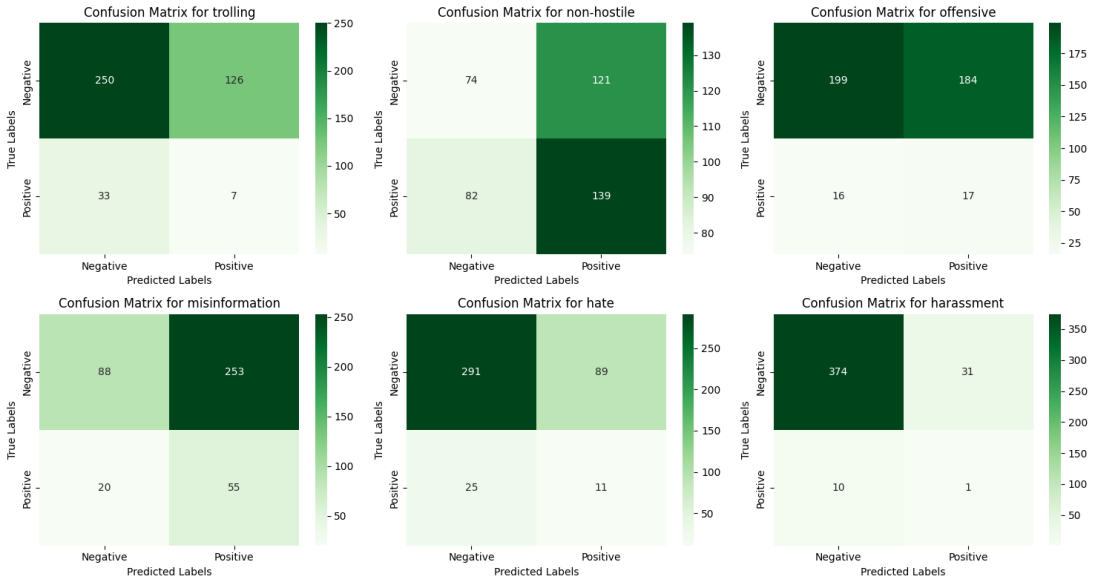


Figure 5.3.3: Confusion Matrix

5.4 Summary

Developing This paper presents a study on the use of Convolutional Neural Networks (CNN) and Bi-LSTM for detecting and classifying hostility in Indonesian Facebook Bangla comments. The method was built using the Keras library for Python and used a dataset with six hostility class labels. The data preprocessing steps involved splitting the text into words, replacing certain patterns with empty strings, removing punctuations, joining words, and applying preprocessing steps.The dataset was split into 80% for training and 20% for validation sets. The experiment was conducted for 1, 2, 3, 4, and 5 epochs. The results showed that the Bi-LSTM model achieved the highest accuracy at the 5th epoch, but the loss increased gradually after 2 epochs. The model checkpoint stopped saving at 5 epochs, achieving the most optimal result at a training accuracy of 87.67%, a training loss of 36.89%. The study also discusses the performance of CNN, Bi-LSTM, and LSTM models in detecting and classifying hostility. It compares the achieved accuracy, loss, val_Acc, val_Loss, and metrics with baseline models or existing approaches. The study also addresses the issue of model interpretability in the context of sensitive hostility and discusses the generalization performance of the CNN Base model.The study also discusses the potential impact of the CNN Base model on content moderation, hostility mitigation, online safety, and fostering healthy online communities. It also highlights the need for ongoing collaboration between AI models and human moderators. The practical applications of the CNN Base model are discussed, with the potential impact on content moderation, hostility mitigation, online safety, and fostering healthy online communities.

CHAPTER 6

Impact on Society, Environment and Sustainability

6.1 Impact on Life

The impact of multi-label hostility classification in Bangla on society can be significant, affecting various stakeholders and aspects of online communication. Here are some potential impacts of our project:

Promoting Online Safety: Effective hostility classification systems can contribute to creating safer online environments by detecting and mitigating harmful content such as hate, harassment, and trolling in Bangla language text. This helps protect individuals and communities from online abuse and discrimination.

Combatting Misinformation and Radicalization: Hostility classification systems can aid in identifying and flagging misinformation, propaganda, and extremist content in Bangla text. By combatting the spread of harmful ideologies and radicalization narratives, these systems help prevent the escalation of social tensions and conflicts.

Enhancing Content Moderation Efforts: Online platforms and social media companies can leverage multi-label hostility classification systems to improve content moderation processes and enforce community guidelines more effectively. This can lead to a reduction in the proliferation of toxic content and improve user experience on digital platforms.

Supporting Law Enforcement and Legal Action: Hostility classification systems can assist law enforcement agencies and legal authorities in monitoring online threats, identifying perpetrators of cyberbullying or online harassment, and gathering evidence for legal proceedings. This can facilitate the enforcement of laws related to cybercrime and hate.

Fostering Inclusive Online Discourse: By identifying and mitigating hostile content, multi-label hostility classification systems promote a more inclusive and respectful online discourse in Bangla-speaking communities. This encourages constructive dialogue, diversity of opinions, and mutual respect among users.

Raising Awareness and Education: The deployment of hostility classification systems raises awareness about the prevalence and impact of online hostility in Bangla text. This can spark conversations about digital literacy, online etiquette, and responsible use of social media, leading to greater public awareness and education on these issues.

Challenges and Ethical Considerations: The implementation of hostility classification systems also raises ethical considerations, such as privacy protection, freedom of expression, and potential biases in algorithmic decision-making. It is essential to address these challenges through transparency, accountability, and stakeholder engagement.

6.2 Impact on Society & Environment

While multi-label hostility classification in Bangla itself may not have a significant direct impact on the environment, it is essential for researchers and practitioners to consider the environmental implications of their computational activities, infrastructure requirements, and data center operations. Implementing energy-efficient practices, utilizing renewable energy sources, and minimizing electronic waste can help mitigate the environmental footprint of hostility classification systems and other AI applications. The impact of multi-label hostility classification in Bangla on the environment is primarily indirect and minimal compared to its social implications. However, there are a few aspects to consider:

Computational Resources: Developing and training machine learning models for multi-label hostility classification requires computational resources, including energy consumption and computing power. While the environmental impact of individual experiments may be relatively small, the cumulative effect of widespread research and deployment of such systems could contribute to overall energy consumption in data centers.

Data Center Operations: Hosting and maintaining servers or cloud infrastructure for deploying hostility classification systems may involve energy-intensive data center operations. However, the environmental impact largely depends on the energy efficiency measures implemented by data center operators and the source of electricity used.

Hardware Production and E-Waste: The production and disposal of hardware components CPUs, GPUs, and other computing devices used for model development and deployment can have environmental implications. Additionally, the disposal of electronic e-waste generated from obsolete or decommissioned hardware can contribute to environmental pollution if not properly managed.

Telecommunication Infrastructure: Transmitting data between users and servers for hostility classification systems may require telecommunications infrastructure, including network infrastructure and data transmission equipment. While the environmental impact of

telecommunications infrastructure is relatively low compared to other sectors, it still contributes to overall energy consumption and carbon emissions.

6.3 Ethical Aspects

Ethical considerations are paramount when developing and deploying multi-label hostility classification systems in Bangla. Here are some key ethical aspects to consider:

Freedom of Expression: Hostility classification systems should not unduly restrict freedom of expression or suppress legitimate speech. It's essential to strike a balance between addressing harmful content and preserving individuals' rights to express opinions and engage in discourse, even if they are controversial or unpopular.

Transparency and Accountability: Transparency about how hostility classification systems work, including their algorithms, training data, and decision-making processes, is essential for building trust and accountability. Users should understand how their content is classified and have recourse if they believe the classification is incorrect or unjust.

Privacy Protection: Hostility classification systems may involve analyzing users' personal data and communications. It's crucial to respect users' privacy rights, protect sensitive information, and adhere to data protection regulations and local privacy laws.

Human Oversight and Intervention: Automated systems for hostility classification may not always capture the complexity and context of human communication accurately. Human oversight and intervention, such as human moderators or appeals processes, are essential for reviewing flagged content, addressing false positives, and making nuanced decisions in ambiguous cases.

Cultural Sensitivity: Hostility classification systems should be culturally sensitive and context-aware, especially when applied to languages like Bangla with diverse linguistic nuances and cultural norms. It's essential to involve linguistic and cultural experts in dataset annotation, model development, and content moderation to avoid misinterpretations and cultural insensitivity.

6.4 Sustainability Plan

Developing a sustainability plan for multi-label hostility classification in Bangla involves ensuring the long-term viability, effectiveness, and ethical use of the classification system

while minimizing environmental impact. Here's a comprehensive sustainability plan:

Hardware Sustainability: Choose energy-efficient hardware components for model development and deployment, and prioritize vendors with sustainable manufacturing practices and commitment to reducing electronic waste. Implement responsible recycling and disposal practices for obsolete hardware to minimize environmental impact.

Data Privacy and Security: Implement robust data privacy measures to protect user data and sensitive information used in hostility classification systems. Adhere to data protection regulations and industry standards for data security, encryption, and access control to mitigate privacy risks and maintain user trust.

Community Engagement: Foster dialogue and collaboration with Bangla-speaking communities, civil society organizations, and advocacy groups to understand their needs, concerns, and priorities related to hostility classification. Solicit feedback, input, and participation in the design, development, and evaluation of classification systems to ensure they meet community expectations and align with cultural sensitivities.

Education and Awareness: Raise awareness about the importance of responsible AI and ethical use of hostility classification systems through education, training, and outreach activities. Provide resources, guidelines, and best practices for users, content creators, and platform operators to promote respectful online communication and combat online hostility effectively.

Continuous Improvement and Adaptation: Regularly evaluate and update hostility classification systems in response to evolving threats, user feedback, and technological advancements. Invest in research and development to enhance the accuracy, fairness, and effectiveness of classification models while minimizing environmental impact and maximizing sustainability.

6.5 Summary

The implementation of multi-label hostility classification in Bangla can significantly impact society and online communication. It can promote online safety by detecting and mitigating harmful content, combating misinformation and radicalization, improving content moderation efforts, supporting law enforcement and legal action, fostering inclusive online discourse, raising awareness and education, and addressing ethical considerations such as privacy protection, freedom of expression, and potential biases in

algorithmic decision-making. The environmental impact of multi-label hostility classification in Bangla is primarily indirect and minimal compared to its social implications. However, it is essential for researchers and practitioners to consider the environmental implications of their computational activities, infrastructure requirements, and data center operations. Implementing energy-efficient practices, utilizing renewable energy sources, and minimizing electronic waste can help mitigate the environmental footprint of hostility classification systems and other AI applications. The environmental impact of multi-label hostility classification in Bangla is primarily indirect and minimal compared to its social implications. However, there are several aspects to consider: computational resources, data center operations, hardware production and e-waste, and telecommunication infrastructure.

Ethical considerations are paramount when developing and deploying multi-label hostility classification systems in Bangla. Freedom of expression should not be unduly restricted, transparency and accountability are essential for building trust and accountability, privacy protection is crucial for preserving users' rights, human oversight and intervention are essential for reviewing flagged content, and cultural sensitivity is crucial for avoiding misinterpretations and cultural insensitivity. Developing a sustainability plan for multi-label hostility classification in Bangla involves choosing energy-efficient hardware components for model development and deployment, prioritizing vendors with sustainable manufacturing practices and commitment to reducing electronic waste, implementing robust data privacy measures, fostering community engagement, raising awareness about responsible AI and ethical use of hostility classification systems, and continually evaluating and updating the system in response to evolving threats, user feedback, and technological advancements. In conclusion, the implementation of multi-label hostility classification in Bangla has significant potential impacts on society and the environment. By considering these factors, researchers and practitioners can develop effective and sustainable solutions to address the challenges and promote inclusivity in online communication.

CHAPTER 7

Conclusion and Future Work

7.1 Conclusions

In conclusion, this study focused on the development and application of Convolutional Neural Networks for hostility detection and classification in the context of social media platforms. The objective was to create an automated system that can accurately identify and categorize 'trolling', 'non-hostile', 'offensive', 'misinformation', 'hate', and 'harassment' 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 hostility. The models achieved high accuracy, loss,val_loss, and val_accuracy, showcasing their potential in automated content moderation systems. We achieved the most optimal result at a training accuracy of 87.67%, and a training loss of 36.89% by using Bi-LSTM mode. The quality and diversity of the training dataset significantly impact the performance of CNN Bi-LSTM and LSTM models. The study identified several challenges and limitations in hostility detection using CNNs. These include the dynamic nature of language, evolving forms of hostility, and the potential for adversarial attacks. Addressing these challenges requires ongoing research and continuous model iteration. The deployment of hostility detection systems based on CNNs can have a positive impact on society. By efficiently identifying and categorizing hostility, 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 hostility 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, we can make significant strides toward combating hostility and promoting a more inclusive and respectful online community

7.2 Further Suggested Works

The study on multi-label hostility classification in Bangla lays the groundwork for further research and exploration in several directions. Here are some implications for further study

in this area:

Fine-grained Hostility Classification: Further research can focus on developing more fine-grained classification models capable of distinguishing between subtle variations of hostility in Bangla text. This could involve refining annotation guidelines, collecting specialized datasets, and exploring advanced model architectures to capture nuanced linguistic cues.

Cross-Lingual Transfer Learning: Investigate the feasibility of cross-lingual transfer learning techniques to leverage labeled data from other languages for hostility classification in Bangla. Transfer learning methods such as multilingual BERT or cross-lingual embeddings could be adapted and evaluated for their effectiveness in this context.

Dynamic Adversarial Detection: Explore techniques for dynamically detecting and adapting to adversarial attacks aimed at evading hostility classification systems. This could involve adversarial training, robust optimization methods, or ensemble approaches to enhance model resilience against adversarial manipulation.

Cultural Sensitivity and Contextual Adaptation: Investigate methods for enhancing the cultural sensitivity and contextual adaptation of hostility classification models for Bangla text. This may involve incorporating socio-cultural knowledge, linguistic variations, and contextual cues specific to Bangla-speaking communities into model development and evaluation.

Interdisciplinary Collaboration: Foster interdisciplinary collaboration between researchers, practitioners, linguists, psychologists, and social scientists to enrich the understanding of online hostility dynamics and develop holistic approaches to addressing hostile behavior in online environments. Collaborative research projects can leverage diverse expertise and perspectives to tackle complex societal challenges.

Ethical Considerations and Human Rights Impact: Conduct in-depth studies on the ethical implications and human rights impact of hostility classification systems in Bangla-speaking communities. This includes investigating issues related to freedom of expression, privacy, bias, discrimination, and the disproportionate impact on marginalized groups.

Real-World Deployment and Evaluation: Explore the challenges and opportunities of deploying hostility classification systems in real-world settings, such as social media platforms, online forums, and news websites catering to Bangla-speaking audiences. Conduct longitudinal studies and field experiments to evaluate the effectiveness,

scalability, and societal impact of deployed systems.

7.3 Limitations

By acknowledging these limitations, our research provides a clear understanding of the constraints and areas for future work, such as expanding the dataset, exploring ensemble methods, and testing on multiple platforms to enhance model robustness and generalizability.

Data Amount: The availability of large, labeled datasets for Bangla text is limited. Training effective deep learning models requires substantial amounts of data, and the lack of sufficient labeled examples can hinder model performance. Besides quantity, the quality of available data is crucial. Incomplete or noisy data can negatively impact the training process and lead to poor generalization.

Single Model Classification: Relying on a single type of model, such as CNN, RNN, or BiLSTM, may not capture all the nuances of hostility in Bangla text. Different models excel at different aspects CNNs are good for feature extraction, RNNs for sequence modeling, and a single model might not be sufficient. Not using ensemble methods or hybrid models can limit performance. Combining multiple models often leads to better accuracy and robustness, as it leverages the strengths of different architectures.

Single Platform: Developing and testing models on data from a single platform only Facebook can introduce platform-specific biases. The language, tone, and type of content can vary significantly across different platforms. Models trained on data from one platform might not generalize well to others. This limitation reduces the applicability and effectiveness of the models in broader contexts or on new platforms.

References

[1] Hossain, M. R., Hoque, M. M., Siddique, N., & Sarker, I. H. (2021). Bengali text document categorization

based on very deep convolution neural network. *Expert Systems with Applications*, *184*, 115394.

[2] Bhardwaj, M., Sundriyal, M., Bedi, M., Akhtar, M. S., & Chakraborty, T. (2023). HostileNet: Multilabel Hostile Post Detection in Hindi. IEEE Transactions on Computational Social Systems.

[3] Elangovan, A., & Sasikala, S. (2022). A Multi-label Classification of Disaster-Related Tweets with Enhanced Word Embedding Ensemble Convolutional Neural Network Model. *Informatica*, *46*(7).

[4] Tripto, N. I., & Ali, M. E. (2018, September). Detecting multilabel sentiment and emotions from banglayoutube comments. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE.

[5] Nafis, N. S. M., & Awang, S. (2020). The evaluation of accuracy performance in an enhanced embedded feature selection for unstructured text classification. *Iraqi Journal of Science*, 3397-3407.

[6] Shakil, M. H. (2022). A hybrid deep learning model and explainable AI-based Bengali hate speech multi-label classification and interpretation (Doctoral dissertation, Brac University).

[7] Naha, S., & Wang, Y. (2016, December). Beyond verbs: Understanding actions in videos with text. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 1833-1838). IEEE.

[8] Sazzed, S. (2020, November). Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020)* (pp. 50-60).

[9] Bhatnagar, V., Kumar, P., Moghili, S., & Bhattacharyya, P. (2021). Divide and conquer: an ensemble approach for hostile post detection in Hindi. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1 (pp. 244-255). Springer International Publishing.

[10] Sinha, A., Naskar, M. N. B., Pandey, M., & Rautaray, S. S. (2022, December). Text classification using machine learning techniques: comparative analysis. In *2022 OITS International Conference on Information Technology (OCIT)* (pp. 102-107). IEEE.

[11] Sarker, M., Hossain, M. F., Liza, F. R., Sakib, S. N., & Al Farooq, A. (2022, February). A machine learning approach to classify anti-social bengali comments on social media. In 2022 International conference on advancement in electrical and electronic engineering (ICAEEE) (pp. 1-6). IEEE.

[12] Banshal, S. K., Das, S., Shammi, S. A., & Chakraborty, N. R. (2023). MONOVAB: An AnnotatedCorpus for Bangla Multi-label Emotion Detection. arXiv preprint arXiv:2309.15670.

[13] Singh, G. V., Priya, P., Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2022). EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. arXiv preprint arXiv:2205.13908.

[14] Das, A., Sharif, O., Hoque, M. M., &Sarker, I. H. (2021). Emotion classification in a resource constrained language using transformer-based approach. arXiv preprint arXiv:2104.08613.

[15] Kabir, A., Roy, A., & Taheri, Z. (2023, December). BEmoLexBERT: A Hybrid Model for Multilabel Textual Emotion Classification in Bangla by Combining Transformers with Lexicon Features. In Proceedings of the First Workshop on Bangla Language Processing (BLP-2023) (pp. 56-61).

[16] Kumar, R., Ratan, S., Singh, S., Nandi, E., Devi, L. N., Bhagat, A., ... & Bansal, A. (2023). A

multilingual, multimodal dataset of aggression and bias: theComMA dataset. Language Resources and Evaluation, 1-81.

[17] Islam, S., Roy, A. C., Arefin, M. S., &Afroz, S. (2022). Multi-label emotion classification of tweets using machine learning. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021* (pp. 705-722). Springer Singapore.

[18] Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2021). Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. Data Science and Engineering, 6(4), 359-379.

[19] Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2020, December). Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In Proceedings of the 28th international conference on computational linguistics (pp. 5810-5820).

[20] Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., & Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework. arXiv preprint arXiv:1910.04602.

[21] Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2020). Fine-grained multi-label sexism classification using semi-supervised learning. In Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21 (pp. 531-547). Springer International Publishing.

[22] Islam, K. I., Kar, S., Islam, M. S., & Amin, M. R. (2021, November). SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3265-3271).

.

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:
MULTI-LABEL HOSTILITY CLASSIFICATION IN BANGLA TEXT

Student ID: 201-15-14119 And 201-15-13834

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a **pneumonia detection** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	Our project demonstrates **fundamental engineering(K3)** principles by employing NLP, data augmentation techniques, and diverse Bi-LSTM architectures for image processing and classification tasks. The project demonstrates **specialist knowledge (K4)** by conducting transfer learning with advanced RNN architectures like Bi-LSTM, and NLP enhancing pneumonia detection accuracy in the Bangla test.	**Page no:** [22-23] **Section:** [3.2] **Page no:** [1] **Section:** [1.1]
				Our project applies **engineering practice & design (K5)** by the figure of the process of experiments. The project addresses **engineering practice & technology (K6)** by employing the Bi-LSTM and CNN model.	**Page no:** [23] **Section:** [3.2(B)] **Page no:** [24-28] **Section:** [3.4]
				Our project ensures **K8 (Research Literature)** by synthesizing insights from recent studies, to Bangla NLP text detection using machine learning, showcasing a comprehensive	**Page no:** [8-18] **Section:**

				understanding of current methodologies.	[2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	Our project addresses**EP-2** by recognizing Bangla raw text data, including the limitations of traditional methods and the complexities of Bi-LSTM and CNNs. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining detection methodologies.	**Page no:** [18-20] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	Our project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting NLP as the chosen significant solution for enhancing Bangla hostility detection.	**Page no:** [31-33] **Section:** [4.2]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting Social hostilitylike 'trolling', 'non-hostile', 'offensive', 'misinformation', 'hate', and 'harassment'to advancements in social media health and protection which indicates **EP-4**.	**Page no:** [8-19] **Section:** [2.2, 2.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting**	No	CO8	N/A	N/A

7.	**EP7: Interdependence**	Yes	CO5	Our project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in Bangla hostility detection which ensures **EP-7**.	**Page no:** [22-29] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, Machine learning frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements in Bangla hostility text data.	**Page no:** [28-29] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving social media through advanced Bangla hostility detection methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for text data privacy.	**Page no:** [36-39] **Section:** [5.1, 5.2, 5.4]

5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in Bangla hostility detection through NLP and CNNs, demonstrated througha comprehensive comparative analysis, offering new insights into the field of Bangla text.	**Page no:** [13-18] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [7-10] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by Bangle people's privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced social communication technologies that comply with **CO9**.	**Page no:** [38] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive Bangla data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [22-29, 31-33] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Hostility

ORIGINALITY REPORT

20%	14%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	aclanthology.org Internet Source	2%
3	doi.org Internet Source	1%
4	Mohit Bhardwaj, Megha Sundriyal, Manjot Bedi, Md Shad Akhtar, Tanmoy Chakraborty. "HostileNet: Multilabel Hostile Post Detection in Hindi", IEEE Transactions on Computational Social Systems, 2024 Publication	1%
5	www.researchgate.net Internet Source	1%
6	link.springer.com Internet Source	1%
7	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%