Cyberbullying and Toxic Language Detection on Social Media for Bangla Language

by

Parom Guha Neogi
20101562
Anjel Haidar Fahim
19101093
Faisal Khan
19101443
Fahim Kabir Khan
19101557
Md. Fahim Faisal
19101161

A thesis submitted to the Department of Computer Science and Engineering
School of Data and Sciences
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

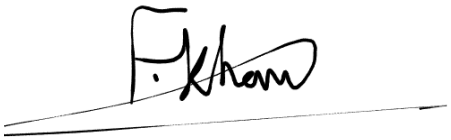
Student's Full Name & Signature:

Parom Guha Neogi

20101562

Anjel Haidar Fahim

19101093



Faisal Khan

19101443

fahim kabir khan

Fahim Kabir Khan

19101557

Md. Fahim faisal

19101161

Approval

The thesis/project titled "Cyberbullying and Toxic Language Detection on Social Media for Bangla Language" Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on submitted by

1. Parom Guha Neogi (20101562)
2. Anjel Haidar Fahim (19101093)
3. Faisal Khan (19101443)
4. Fahim Kabir Khan (19101557)
5. MD. Fahim Faisal (19101161)

Examining Committee:

Supervisor (Member):

Najeefa Nikhat Choudhury

Lecturer

Department of Computer Science and Engineering
Brac University

Co-supervisor (Member):

Md Faisal Ahmed

Lecturer

Department of Computer Science and Engineering
Brac University

Co-supervisor 2 (Member):

Md. Moynul Asik Moni

C. Lecturer

Department of Computer Science and Engineering
Brac University

Head of Department (Chair):

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor

Department of Computer Science and Engineering
Brac University

Abstract

As more people use social media, toxic language and cyberbullying become more common with the Bengali-speaking community particularly. The complexity of Bangla text data makes it difficult for traditional natural language processing (NLP) algorithms to identify harmful content. This study proposes a machine learning-based solution that recognizes and categorizes harmful language and "Cyberbullying in Bangla text on social media", leveraging BanglaBERT's advanced features. As more people use social media, toxic language and cyberbullying are on the rise, with the Bengali-speaking minority particularly vulnerable. The proposed machine learning-based solution achieved 94% testing accuracy in detecting and categorizing cyberbullying and offensive language on digital platforms that support Bengali.

Keywords: Cyberbullying; Detection; Internet; Language; BERT, NLP

Dedication

We have the utmost respect for the Almighty and devote this work of ours to Him.

We want to offer our sincere appreciation to each and every member of the group T2310090 for their incredible work and steadfast support throughout our difficult research.

We will always be grateful to our parents for their continuous support and belief in our abilities, which helped us attain academic success. They helped us progress and mold into the people we are today with their uplifting words and sincere prayers. We recognize and appreciate the immense worth of their self-sacrifice as we get ready to embark on the next chapter of our lives. We sincerely appreciate their unwavering support.

Acknowledgement

We start by expressing the Almighty our sincere gratitude and reverence, as his unwavering assistance and favor were essential to the effective conclusion of our thesis. His spiritual counsel illuminated our way, gave us the strength, and informed us enough to overcome the major challenges in our academic journey.

We would like to appreciate Najeefa Nikhat Choudhury, our well-respected supervisor, and Md. Faisal Ahmed, our devoted co-supervisor, for their crucial efforts to make our job more understandable. They functioned as role models, helping us overcome challenges and providing us the self-assurance to deal with problems with their knowledge, astute advice, and unwavering support.

We also appreciate their willingness and availability to assist us whenever we needed it. We were able to refine our theories and procedures as a result of their attentive listening, adaptive analysis, and quick meditations, which raised the likelihood of our thesis.They created an environment that encouraged growth, support, and learning, and their tutoring extended beyond the classroom. Because of their kind observations and encouraging actions, we had the will and courage to endure in the face of uncertainty and exhaustion.

We truly thank them for the tremendous time, effort, and expertise they invested in developing our academic progress, despite their demanding duties. Academically, their ongoing commitment to our prosperity has a significant impact on our lives now and in the future.

Table of Contents

List of Figures

List of Tables

Nomenclature

Several symbols and abbreviations that will be used later in the document's body are described in the following list.

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Network

ML Machine Learning

NLP Natural Language Processing

RF Random Forest

RNN Recurrent Neural Network

SVM Support Vector Machine

XLM-ROBERTa Cross-lingual Language Model - Robustly Optimized BERT Approach

Chapter 1

Introduction

For everyday communication, we use a number of social media platforms, including Facebook, Twitter, WhatsApp, and Instagram. Some of us write and publish blog posts on our own websites. Sadly, an increasing number of people are using technology to send offensive or threatening messages to other people. Cyberbullying is the term used to describe this action. Social media allows for rapid information access, which might worsen the issue. Online abuse is a dangerous and harmful activity that could affect anyone. Although social media platforms have the benefit of facilitating user communication and information exchange. Yet, they have enabled harmful behaviors such as online harassment, spreading false information, and hate speech. Hate crimes may result from this sort of communication, which often targets certain populations. It frequently emerges on social networks and other internet platforms. According to reports, in January 2023[1], more than 44.70 million users, or over 26% of Bangladesh's population, accessed social media. Over the same time frame, there were 179.9 million active mobile connections, or nearly 104.6% of the total population. Facebook is used by more than 90% of Bangladesh's social media users, over 4.9 million Instagram users of the total population and nearly everyone who owns a smartphone and access to the internet uses WhatsApp, with the majority of who are young and impulsive. People who are impulsive are more likely to engage in cyberbullying because they do not carefully consider the consequences of their actions. Out of rage or frustration, they may act cruelly toward others, utter hateful words, respond to perceived slights, or take some other action.

In today's world, languages are actively evolving, with communities influenced by an ever-changing world. Such massive amounts of data necessitate manual inspection and verification, which is labor-intensive as well as time-consuming and prone to human error. Hate speech on social media must be detected and addressed, especially in languages such as Bengali, which has 230 million speakers but is significantly underfunded in terms of natural language processing (NLP). This is owing to a scarcity of NLP language models, tagged datasets and effective machine learning algorithms.

1.1 Research Problem

While implementing a machine learning-based approach for detecting and preventing cyberbullying, there are several challenges that researchers may face. We will discuss some of the potential problems that may arise during the project.

1.1.1 Data Collection and Labeling

Obtaining a sufficient amount of labeled data for cyberbullying detection was challenging. Cyberbullying incidents were sporadic and dispersed across various online platforms, making it difficult to collect comprehensive data. When we started collecting data, we found only sexual and troll types of data; there was not enough data on religion and threat. We didn't even get large amounts of data due to social media terms. Sometimes we encountered the same data, but labeling was different. We worked on some recent social media activities, where in the comment boxes we found repetition of the same comments. We had to struggle to collect religious and threat comments because social media policies were stronger than before. Sexual and troll comments were available on platforms like Facebook, but other types were scarce.

1.1.2 Bias in Data

When we collected or analyzed the data with bias, it indicated that the data was not entirely perfect. We gathered data from various social media platforms, and sometimes the same data was presented from different viewpoints. This discrepancy could occur when certain groups were favored over others in the data collection process, potentially leading to misleading results. Therefore, we had to employ different methods depending on how the data was collected.

If the training data predominantly consisted of examples from certain demographics or cultural backgrounds, the model might have struggled to accurately detect cyberbullying directed towards under-represented minority groups[2]. This could result in biased predictions, where instances of cyberbullying against these groups were overlooked or misclassified. Bias could also arise from subjective labeling processes or from stereotypes and prejudices influencing the labeling of examples in the training data.

Moreover, the language used in online communication platforms often contains inherent biases or cultural nuances that might not have been adequately represented in the training data. This could pose challenges for the model in accurately identifying instances of cyberbullying that deviated from the language patterns present in the training set. Additionally, biases could be introduced during the process of collecting training data if certain sources or platforms were over-represented while others were under-represented. This could skew the model's perception of what constitutes cyberbullying, leading to biased predictions.

Furthermore, machine learning algorithms themselves could introduce bias if they were designed or trained in a manner that disproportionately affected certain groups. For instance, if the features used by the model to detect cyberbullying were biased towards a particular demographic, the model's predictions might reflect this bias. Therefore, mitigating bias in both the data collection process and the algorithm

design was crucial for developing fair and effective cyberbullying detection systems.

1.1.3 Overfitting

Despite having a sufficient amount of data in our database, we encountered difficulties in obtaining accurate results when incorporating new data. Our reliance on a single sample set of data often consumed most of our time. Sometimes, the size of our data was insufficient to yield precise outcomes. An overfitted model, which arises from excessive tuning to the training data, may erroneously classify instances as cyberbullying when they are not, due to capturing noise or irrelevant patterns present in the training data. This can lead to false accusations and unnecessary interventions, potentially harming innocent individuals.

Overfitting leads to a model that is overly specialized to the training data, diminishing its effectiveness at identifying cyberbullying in new, unseen data[3]. Consequently, the utility of the detection system may be undermined in real-world scenarios where the data distribution may vary from the training set. When overfitting occurs, the model's performance metrics such as accuracy, precision, and recall may appear overly optimistic when evaluated on the training data. However, these metrics may not accurately reflect the model's performance on new data, leading to a misleading assessment of the system's effectiveness.

Moreover, overfitted models may struggle to adapt to changes in online behavior or emerging forms of cyberbullying, as they become too rigidly tied to the idiosyncrasies of the training data. This inflexibility can render the detection system obsolete over time as cyberbullying trends evolve and grow. Thus, mitigating overfitting and ensuring the model's robustness to new data are critical for maintaining the efficacy of the cyberbullying detection system in dynamic online environments.

1.1.4 Generalization

When focusing solely on Bengali language cyberbullying, we encountered significant challenges due to the diversified nature of the data, which included both English and Bangla content. Consequently, we struggled to achieve accurate outcomes when incorporating new data. Generalization, in this context, refers to the model's ability to perform well on unseen or new data beyond the training data it was exposed to. It entails learning patterns and relationships from training data and applying them accurately to new, previously unseen examples.

However, if the training data used to develop the model is not representative of the broader population or contains biases, the model may fail to generalize well to new data. This could result in the model being overly sensitive to certain types of cyberbullying or failing to detect emerging forms of cyberbullying that were not adequately represented in the training data. Additionally, an overfitted model may perform well on the training data but struggle to generalize to new instances of cyberbullying, leading to poor performance on real-world data.

Furthermore, mismatches between the distribution of the training data and the distribution of the data encountered in the real world can also hinder generalization. Changes in online behavior over time or differences between online platforms could lead to domain shift, causing the model's performance to degrade over time. Hence, ensuring that the model can generalize effectively to new data is essential for the accurate detection of cyberbullying, especially when dealing with diverse linguistic contexts like Bengali language cyberbullying.

1.1.5 Ethical Concerns

Ensuring the privacy and security of participants' identities is paramount, particularly considering the potential trauma associated with cyberbullying. Biases in the training data towards specific types of bullying may result in unjust outcomes, necessitating careful consideration in data collection and analysis to mitigate such biases. However, gathering and analyzing data must not infringe upon individuals' privacy rights, especially if sensitive information is obtained without their consent.

Transparency and accountability are crucial for cyberbullying detection systems to gain trust and ensure fairness. Lack of transparency can lead to mistrust, while accountability is essential for addressing concerns and rectifying errors. Nonetheless, there exists a delicate balance between recognizing harmful behavior and safeguarding free expression. Overzealous identification technologies might inadvertently suppress valid speech, raising concerns about censorship.

Moreover, the accuracy and effectiveness of detection systems raise ethical concerns. False positives can unjustly accuse innocent individuals, while false negatives can allow harmful behavior to persist undetected. Both scenarios have significant ethical implications, emphasizing the need for rigorous testing and continuous improvement to minimize errors and ensure fairness in cyberbullying detection.

1.1.6 Interpretability

Interpretability was crucial in our cyberbullying detection systems to understand how the model made decisions and to ensure transparency and accountability. However, we encountered challenges in obtaining accurate information when entering comments containing different symbols or local slangs unfamiliar to the model. This hindered the model's ability to provide proper responses, especially when dealing with large volumes of data requiring instant feedback.

As the volume of online communication data continued to grow, scalability remained essential for our cyberbullying detection systems. We faced challenges when the algorithms or infrastructure used could not efficiently process and analyze large datasets without experiencing performance bottlenecks or resource constraints. In situations where real-time cyberbullying detection was essential, scalability became even more critical, necessitating the ability to analyze incoming data streams rapidly and make

timely predictions without introducing significant latency. However, the complexity of machine learning models designed to capture intricate patterns in online communication data often hindered scalability, especially when dealing with large-scale datasets or high-dimensional feature spaces.

Adaptability was also key for our cyberbullying detection systems to remain effective amidst evolving online platforms, communication channels, and forms of cyberbullying. We encountered challenges when the system's architecture or algorithms were not flexible enough to accommodate changes in data sources or user behavior without requiring significant re-engineering or retraining. Therefore, ensuring Interpretability, scalability, and adaptability were crucial considerations in the development and deployment of our cyberbullying detection systems.

1.1.7 Scalability

As the volume of online communication data continued to grow exponentially, our cyberbullying detection systems needed to efficiently process and analyze large datasets. We faced scalability challenges when the algorithms or infrastructure used in the detection system couldn't handle the increasing volume of data without experiencing performance bottlenecks or resource constraints.

In cases where real-time cyberbullying detection was essential, scalability became even more critical. The detection system had to analyze incoming data streams rapidly and make timely predictions without introducing significant latency. We encountered scalability challenges when the system couldn't scale up or down dynamically to accommodate fluctuations in workload or traffic.

Complex machine learning models designed to capture intricate patterns in online communication data often proved less scalable than simpler models. As the complexity of the model increased, so did the computational cost of training and inference, potentially leading to scalability issues, especially when dealing with large-scale datasets or high-dimensional feature spaces.

Our cyberbullying detection systems needed to be adaptable to new online platforms, communication channels, and evolving forms of cyberbullying. However, scalability challenges arose if the system's architecture or algorithms weren't flexible enough to accommodate changes in data sources or user behavior without requiring significant re-engineering or retraining efforts. Therefore, ensuring scalability, adaptability, and flexibility were crucial considerations in the development and deployment of our cyberbullying detection systems.

1.2 Research Objective

The main goal of this dissertation is to develop, implement, and assess a comprehensive system that can detect instances of cyberbullying in social media posts, with a particular focus on Bengali-language content. The inquiry seeks to address the

wide range of cyberbullying behaviors, from overt hostility to more subdued forms of harassment, all while navigating the special linguistic difficulties and dearth of Bengali language resources. The dissertation is structured around multiple main goals:

Create a Framework for Linguistically Sensitive Detection: The aim is to build a framework that can identify the unique linguistic, cultural, and contextual quirks that are peculiar to Bengali. This will entail compiling or improving datasets labeled with examples of cyberbullying in Bengali and adapting machine learning algorithms to overcome the unique challenges of the language.

Make Use of Cutting-Edge Machine Learning Innovations: In order to identify cyberbullying, this goal comprises researching and utilizing the most recent developments in machine learning, with a focus on deep learning and transformers. In order to assess these contemporary approaches' effectiveness and suitability for the Bengali environment, a comparison with traditional machine learning techniques will also be made.

Comprehensive Assessment of System Performance: The study will carefully assess how well the created system can identify instances of cyberbullying in social media posts written in Bengali. Metrics like the system's recall, precision, and ability to discern between benign content and cyberbullying in a variety of scenarios and cyberbullying manifestations will be given priority during the evaluation.

Improve Awareness of Cyberbullying and Preventive Measures: In addition to improving digital safety and inclusivity, this dissertation aims to increase the awareness of cyberbullying among Bengali-speaking social media users. It goes beyond the technical accomplishments of developing a detection system.

By focusing on the Bengali language, this dissertation aims to fill a significant gap in the literature on cyberbullying and improve our understanding and detection techniques in less researched linguistic and cultural contexts. It is anticipated that the research will increase the technical defenses against cyberbullying and encourage a safer and more polite online environment for people who speak Bengali.

Chapter 2

Literature Review

According to a research study of Yadav et al. cyberbullying is one of the major problems in social media platforms[4]. The research study claims that the BERT model, which makes use of deep neural networks, significantly improves cyberbullying detection, with 98% accuracy on the Formspring and Wikipedia datasets. Since 2018, Google has integrated BERT into its search engines by using datasets from Wikipedia and Formspring. BERT works well with bigger datasets, therefore Wikipedia does not require any oversampling. It is advised to use Google Colab and Python for implementation. With a significant improvement over earlier models, Formspring accuracy now stands at 98% instead of 78.5%.

A study by Behzadi et al. says, the majority of individuals use social media platforms to promote hate speech, and this trend is growing daily[5]. As a result, we need to discover a solution that can identify cyberbullying more effectively and quickly. To tackle the class imbalance in the dataset, they paired the focal loss function with several bidirectional encoder representations from Transformer or BERT models that were fine-tuned over the hate speech dataset. Besides, BERT has a drawback to its original model due to the large networks and so many transfer layers and hidden embeddings and fine-tuning them with smaller datasets does not provide any best results. They have used Focal Loss in order to solve this situation. Also, they have used FL Hyper-parameter Decision where 90% of data were randomly split and 10% of data was used for validation purposes. They performed 10-fold cross-validation on the complete dataset to assess performance and compare their final findings to earlier work by Founta et al. Along with Founta et al, they utilized many BERT models, including BERT-Base, BERT-Medium, BERT-Small, BERT-Mini, and BERT-Tiny. All of these compact BERT models gave better results than the Founta et al model which had only 84% accuracy level, whereas compact BERT models had around 91% accuracy and the BERT-Base model provided the best result with a 91.56% accuracy level compared to all other compact BERT models.

According to a research study by Gada et al. detection of cyberbullying has become one of the biggest issues in today's society, and with the explosive expansion of social media users, the incidence of cyberbullying is also surging at an alarming pace[6]. This body of literature focuses on using eXtreme Gradient Boosting (XGBoost), Random Forest and Logistic Regression to identify cyberbullying and comparing the outcomes with those of LSTM-CNN models, which are basically Deep Neural

Network based models. Even if there have historically been a lot more machine learning algorithms utilized, the accuracy may always be increased. At first, it was used on tweets to identify the toxicity level and based on that it classified them accordingly. The experiment included visualization of statistics and search methods to find cyberbullying. Later it was implemented on Web Applications, Telegram Bot and Google Chrome Extensions to detect cyberbullying. All incoming text messages and images were run through the system during the implementation process across multiple platforms to determine whether or not cyberbullying had occurred. After the implementation of all the models, we get a clear vision regarding the results. With scores of 38% and 51%, respectively, XGBoost is the best performer in terms of Recall and F1 Score, while Random Forest has the greatest Precision score of 80%. And LSTM-CNN gets the highest Accuracy among all other models with a score of 95.2%.

Raj et al.'s work shows that hybrid models that include machine learning and natural language processing perform better than traditional methods for detecting cyberbullying[7]. In addition to demonstrating robustness against imbalanced datasets, these models get superior F1-scores, recall, accuracy, and precision. Its narrow emphasis on just two cyberbullying datasets, however, may restrict the study's generalizability. It may also provide difficulties in real-world applications since training requires a large amount of labelled data. To improve the study, it could be helpful to investigate semi-supervised techniques and broaden the dataset to encompass a variety of cyberbullying scenarios. Numerous machine learning approaches, including decision trees, support vector machines, random forests, convolutional neural networks, LSTM networks, and attention mechanisms, are included into the hybrid models that are being studied. Reported accuracy rates range from 85% to 95%, indicating the effectiveness of these models in accurately identifying cyberbullying.

The study by Haidar et al. focused on developing a multilingual system for detecting cyberbullying, particularly in Arabic content[8]. The authors aim to address the growing concern of cyberbullying in Arab countries and propose a solution that can detect such behavior in online content . Naive Bayes and SVM were two machine learning methods that the researchers utilized to train their system.. These algorithms were chosen based on previous research in the field of text classification. The system was trained multiple times to achieve the best results. The Naïve Bayes algorithm was found to be more promising in detecting cyberbullying in Arabic content. The authors conducted experiments to evaluate the accuracy of their system. They used a dataset of 1,000 Arabic tweets that were manually labeled as either cyberbullying or non-cyberbullying. The system achieved an accuracy rate of 87%, which is considered high for this type of task. One drawback of this research is that it only focuses on detecting cyberbullying in Arabic content. It would be beneficial to expand this system to other languages as well since cyberbullying is not limited to one language or culture. Additionally, the dataset used for evaluation was relatively small and may not be representative of all types of online content. To improve this research, the authors could consider using larger datasets that include different types of online content such as comments on social media platforms or blog posts. They might also look into different machine learning methods or algorithms that might raise the accuracy rate even higher. In summary, this study offers a practi-

cal method for identifying cyberbullying in Arabic material using machine learning techniques like Naive Bayes and SVM. The system achieved an accuracy rate of 87% when evaluated on a dataset of 1,000 Arabic tweets manually labeled as either cyberbullying or non-cyberbullying. However, there is room for improvement by expanding the system to other languages and using larger datasets for evaluation.

The study by Hani et al. demonstrated promising results in combating cyberbullying through supervised machine learning techniques[9]. Neural Network classifiers achieved the highest accuracy of 92.8%, outperforming Random Forest and Support Vector Machines. Comparison with similar studies revealed the superiority of their approach. However, the study's limitation lies in its exclusive focus on language patterns, neglecting other mediums like images or videos. To enhance their method, integrating image/video analysis and diversifying the dataset beyond language patterns is recommended. Despite this drawback, the study underscores the potential of machine learning in identifying and preventing cyberbullying instances, albeit with room for further refinement.

The study by Chatzakou et al. focuses on detecting abusive behavior on Twitter using a combination of algorithms and models[10]. They achieved an accuracy rate of 80% using Support Vector Machines and Naive Bayes classifiers. However, the study has limitations, such as focusing solely on Twitter data and relying heavily on keyword-based searches. Future research could expand to other social media platforms and employ more advanced natural language processing techniques. Despite limitations, the study offers valuable insights into detecting cyberbullying and cyber aggression, laying a foundation for further advancements in this field.

The analysis of Akhter et al. aims to address the growing concern of cyber bullying and proposes a method for detecting and classifying it using advanced techniques[11].The researchers used a combination of Multinomial Naïve Bayes and Fuzzy Logic algorithms to develop their model. They conducted experiments on a dataset consisting of 10,000 tweets collected from Twitter, which were manually labeled as either cyber bullying or non-cyber bullying. The researchers then used this dataset to train their model and evaluate its performance. The accuracy rate achieved by the proposed model was 87%. Overall, this paper provides valuable insights into how advanced techniques can be used to detect and classify cyber bullying in social media data.

The research study by Ahmed et al. on cyberbullying detection using deep neural networks is found to be useful in handling online harassment in Bengali-speaking communities[12]. They developed a cyberbullying detection model for Bengali-speaking communities using a hybrid neural network trained on a dataset of 44,001 comments from Facebook. The model categorized comments into five categories: non-bully, sexual, threat, troll, and religious. They utilized binary and multiclass classification models with word embedding for representation. The ensemble approach was also employed. Although the model showed promising results, achieving an 87.91% accuracy in binary classification, there are limitations in its generalizability and scalability. The authors suggested improving training time, reducing false positives, and enhancing performance using more sophisticated deep learning

methods like CNNs or RNNs. In order to improve the accuracy of cyberbullying identification, they also suggested expanding the dataset to include comments from a variety of online platforms and languages. They also suggested looking into other variables like user data or network architecture.

The study conducted by Hussain et al. aimed to develop an algorithm capable of identifying offensive Bangla text, distinguishing between categories such as amusing, inflammatory, furious, and abusive phrases[13]. Utilizing a dataset of 500 Bangla comments collected from various social media platforms, they employed machine learning methods including Naive Bayes, Support Vector Machine, and Random Forest. Their suggested method achieved an accuracy rate of 85%. However, it solely focused on detecting abusive Bangla text and lacked consideration for other languages. Furthermore, the algorithm's real-time applicability is limited due to preprocessing requirements. Expanding the dataset to include additional comments from other platforms and investigating alternative machine learning methods like convolutional neural networks or recurrent neural networks are two recommendations for improvement. Even with encouraging outcomes, more work has to be done to improve precision and allow for real-time detection, which is important for the fight against cybercrime in Bangladesh.

In the study of Talpur et al. the application of machine learning algorithms for the identification of cyberbullying has been the subject of a large amount of study in recent years[14]. The researchers used a dataset of tweets containing cyberbullying content, which they manually labeled as either cyberbullying or non-cyberbullying. Then, they extracted characteristics from the tweets using a variety of NLP approaches, including sentiment analysis and part-of-speech tagging.The labeled dataset was then used by the researchers to train a number of machine learning models, including logistic regression, decision trees, and support vector machines (SVMs). They discovered that SVMs had the highest accuracy, with an F1 score of 87

In this research study Sarker et al. defined an unique method for identifying false news using machine learning techniques is presented[15]. This strategy's value resides in its capacity to stop the spread of false information and raise media literacy levels among the general populace. It does not provide specific information on the algorithm used in the study. However, it mentions that the researchers developed a model using machine learning techniques and various features to predict whether an article is likely to be fake or not. The researchers developed a model that uses various features, such as the source of the news article and the language used, to predict whether an article is likely to be fake or not. The model has an accuracy rate of almost 90% after being trained and evaluated on a big dataset of news stories. However, there were some drawbacks to the study. One limitation was that the dataset used for training and testing may not be representative of all types of fake news. Additionally, there may be other factors that contribute to the spread of fake news that were not included in the model. To improve upon this research, future studies could incorporate more diverse datasets and consider additional variables that may impact the spread of fake news. Overall, the accuracy rate achieved by this model indicates its potential as a useful tool for detecting and combating fake news in today's media landscape.

Alsaed et al. in their study, thoroughly examined the methods currently in use for identifying cyberbullying on social networking sites[16].. These methodologies were divided into three categories by them: hybrid tactics, unsupervised learning algorithms, and supervised machine learning techniques. Although the study did not present any novel detection techniques, it performed a thorough analysis of ten years of relevant research. The work is divided into parts that address performance comparisons, definitions of cyberbullying, detection methods, open research questions, and future goals. Future research might focus on creating algorithms that are more precise and effective, even if the study's primary objective was to examine current methodologies. Even though the assessed articles' accuracy rates varied, the majority claimed high rates ranging from 80% to 95%.Overall, the review underscores the importance of current detection methods and suggests avenues for improvement to combat cyberbullying effectively on social media platforms.

The study by Ghosh et al. on cyberbullying detection in Bengali language reveals promising results using machine learning techniques[17]. Random Forest classifier achieved an accuracy of 60.3% with word-level features and 77.1% with N-gram-level TF-IDF data.Support Vector Machine classifier achieved 61.1% accuracy with word-level features and 77.7% with N-gram features.Logistic Regression yielded accuracy rates of 58.8% and 77.5% for word and N-gram-level features, respectively. Passive-aggressive classifier achieved the highest accuracy of 78.1% with N-gram-level TF-IDF data but only 51.0% with word-level data.

According to Emon et al. the identification of hate speech and cyberbullying through internet entertainment has become one of the most crucial language control strategies in recent years[18]. However, due to a lack of resources, detecting cyberbullying in Bangla has received less attention. This paper attempts to address this need by recognizing cyberbullying in Bangla virtual entertainment through the use of a few transformer models. We use a dataset of 44,001 Bangla comments grouped into five classes based on Facebook postings: sexual, destructive, strict, and savage. Cyberbullying is recognized using three transformer models: Bangla BERT, Bengali DistilBERT, and XLM-RoBERTa. These transformer models are intricate language models that are capable of encoding both vast amounts of data and semantic meaning. These methods are used in the review to provide accurate cyberbullying detection in Bangla. The dataset, which contains measurements like accuracy and F1-score, is used to create and test transformer models. According to the test results, the transformer models are astonishingly good at spotting cyberbullying in Bangla social media. With an F1-score of 86% and accuracy of 85%, the XLM-RoBERTa model performs better than the competitors. The accuracy and F1-score obtained reflect how well the algorithms were able to identify and classify occurrences of cyberbullying in the social media data from Bangladesh.

The research study of Sultana et al. states that virtual entertainment has many benefits, but the abundance of unfavorable remarks has turned into a challenging problem. To solve this problem, researchers searched for and analyzed critical Bengali comments posted on online entertainment sites using machine learning techniques. Although Bengali has garnered a lot of attention in other languages, there aren't many studies that focus solely on it. The efficacy of machine learning computations

for identifying disagreeable statements was evaluated using K-Nearest Neighbor, Random Forest, Support Vector Machine, Logistic Regression, and Gradient Boosting. It was crucial how well these algorithms fixed problems with text order. The dataset used in this study consists of Bengali-language comments that are separated.For finding objectionable remarks in Bengali virtual entertainment, according to preliminary statistics Support Vector Machine performed the best. The accuracy of 85.7% of the SVM calculation not only demonstrates how well it can recognize and categorize negative articulations in Bengali but also how effectively it can handle Bengali's distinctive linguistic depth and use patterns.

Comparison between work on Bengali Language			
Author	**Focus**	**Methods**	**Performance**
Hussain et al.	Offensive Text Detection in Bangla	Naive Bayes, SVM, Random Forest	Accuracy: 85%
Emon et al.	Cyberbullying Detection in Bangla	Transformer Models (Bangla BERT, Bengali DistilBERT, XLM-RoBERTa)	Accuracy: 85%
Sultana et al.	Offensive Phrase Recognition in Bengali	Support Vector Machine (SVM)	Accuracy: 85.7%
Haidar et al.	Multilingual Cyberbullying Detection	Naive Bayes, SVM	Accuracy: 87%
Ahmed et al.	Cyberbullying Detection in Bengali	Hybrid Neural Network, Ensemble Approach, Word Embedding	Accuracy: 87.91%
Behzadi et al.	Hate Speech Detection using compact BERT	Compact BERT	Accuracy: 91.56%
Hani et al.	ML Techniques for Cyberbullying Detection	Neural Network Classifiers (NNC)	Accuracy: 92.8%

Table 2.1: Comparison between the Models

Chapter 3

Dataset Description

Cyberbullying is becoming more common in today's digital age, especially on social media sites. It involves hurting, forcing, or mistreating people using digital technology. Because of the large amount of text people post every day, it can be hard to spot cyberbullying. Researchers have been looking into using natural language processing (NLP) to automatically detect cyberbullying on these platforms.

3.1 Data Collection

The initial dataset utilized in this study, comprising 44,001 comments, was meticulously compiled and made publicly available on Mendeley Data under the title "Bangla Online Comments Dataset". This dataset contains a varied range of interactions from many social spheres, such as comments made by actors, influencers, politicians, and sportsmen, illustrating the diverse terrain of online discourse in Bangla language. To guarantee completeness and representativeness, 16,000 extra data points were collected from sites such as Facebook, Instagram and YouTube[19]. Rigorous peer validation processes were employed to maintain the integrity and accuracy of the dataset labels, resulting in a reliable resource for further analysis and research endeavors. The dataset includes comments classified as sexual, non-bully, troll, religious, and threat, offering insight into various aspects of online communication in the Bangla-speaking population.

	comment	Category	Gender	comment react number	label
0	ওই হ্যালার পুত এখন কি মদ খাওয়ার সময় রাতের বেলা...	Actor	Female	1.0	sexual
1	ঘরে বসে শুট করতে কেমন লেগেছে? ক্যামেরাতে কে ছি...	Singer	Male	2.0	not bully
2	অরে বাবা, এই টা কোন প্যাগল????	Actor	Female	2.0	not bully
3	ক্যাপ্টেন অফ বাংলাদেশ	Sports	Male	0.0	not bully
4	পটকা মাছ	Politician	Male	0.0	troll
5	অন্যরকম .. ভালো লাগলো ..❤️	Singer	Male	1.0	not bully
6	সাংবাদিক ভাইদের বলছি এই সংবাদ গুলি প্রচার না ক...	Actor	Female	9.0	troll
7	মোহাম্মদ কফিল উদ্দীন মাহমুদRidwan RomelDwaipay...	Actor	Female	0.0	not bully
8	ঢাকায় এত ঘনো ঘনো আগুন লাগার মূল কারন টা এতদিনে...	Actor	Female	4.0	religious
9	হিরো আলম তুমি এগিয়ে চলো, আমরা আছি তোমার সাথে।	Social	Male	0.0	not bully

Figure 3.1: Original Dataset

Initial Dataset Compilation: An initial dataset of 44,001 comments was used in the study. The research's basis was created by carefully gathering and compiling these comments. To provide transparency and accessibility for other researchers, the dataset was made publicly available on Mendeley Data.

Source of Comments: The comments have been gathered from a variety of internet sources, with a focus on posts on Facebook that were open to the public. This implies concentrating on a well-liked and extensively used social media site in order to record a large variety of interactions.

Diverse Interactions: The dataset includes comments from a wide range of individuals and entities, such as actors, social media influencers, singers, politicians, and athletes. This diversity reflects the broad landscape of online discourse, encompassing different topics and perspectives.

Expansion of Dataset: To enhance the comprehensiveness and representativeness of the dataset, an additional 16,000 data points were collected from various social media platforms, including Instagram, YouTube, and others. This expansion aims to capture a broader spectrum of interactions across different online communities.

Category	Number of Comments
Sexual	10,046
Non-bully	19,269
Troll	14,702
Religious	8,689
Threats	5,742
Total	58,448

Table 3.1: Comments by Category

Manual Collection: To ensure a thorough perspective of online debate, extra data points were collected by human sourcing. This hands-on method most likely required human judgement and selection to gather meaningful remarks from a variety of sources.

Data Labeling and Validation: The integrity and quality of dataset labelling were critical. A robust peer validation procedure was created, with the team cross-validating the given labels for each remark. This method required thorough analysis to assure accuracy and dependability, with any errors or ambiguities in labelling resolved through joint efforts.

Final Dataset Composition: The resulting dataset comprises a total of 58,448 comments extracted from various online platforms. Each comment is categorized into

specific types, including sexual, non-bully, troll, religious, and threats, providing a detailed breakdown of the dataset's composition.

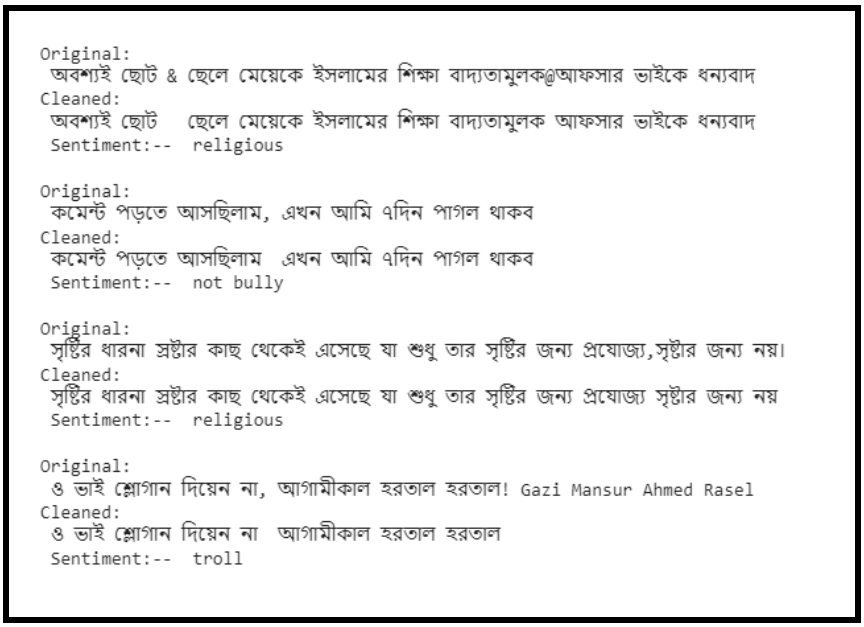


Figure 3.2: Final Dataset (Cleaned)

3.2 Data Analysis

3.2.1 Gender

The gender distribution within the dataset, with approximately 36,000 comments associated with female individuals and the remaining comments pertaining to male individuals, offers a nuanced perspective on online discourse. This meticulously curated and validated dataset serves as a valuable resource for researchers and analysts interested in understanding the intricacies of online communication.

The dataset's comprehensiveness means that it covers a wide range of topics and circumstances, giving valuable insights into numerous elements of online interactions. By thoroughly curating the dataset, researchers may have trust in the data's accuracy and dependability, allowing them to conduct robust analyses with confidence.

Furthermore, the comprehensive validation method used to create this dataset improves its legitimacy and value. By verifying the dataset's comments, researchers may confirm that the data truly reflects real-world online interactions, reducing biases and mistakes.

This dataset has the potential to offer insight on topics such as harassment, discrimination, and social dynamics in online places. Researchers may use this information to investigate behavioural patterns, detect trends, and reveal underlying variables driving online communication dynamics.

Furthermore, the insights gleaned from analyzing this dataset can contribute to a deeper understanding of societal trends and attitudes towards gender within online communities. By examining how different genders are represented and treated in online discourse, researchers can gain valuable insights into broader social dynamics and inequalities.

Finally, this thoroughly curated and validated dataset is an invaluable resource for researchers exploring the complexity of online communication. Its extensive scope, rigorous validation procedure, and emphasis on gender distribution make it an effective tool for better understanding online controversy and social trends.

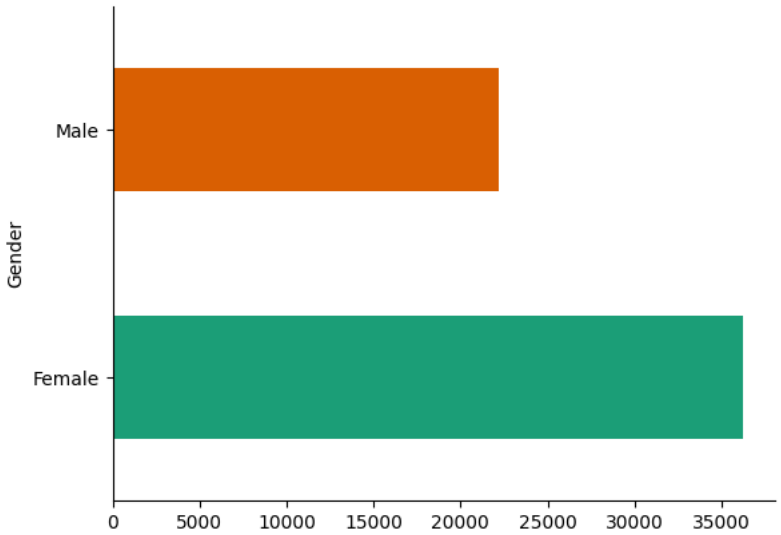


Figure 3.3: Gender Distribution of the Dataset

3.2.2 Category distribution by profession

The dataset gives an in-depth overview of the prevalence of online harassment across numerous occupational categories, putting light on the problem's prevalence in today's digital society. Let's go deeper into the significance of each profession mentioned in the analysis:

Acting Industry (33,067 comments): With a significant amount of comments aimed at actors and actresses, the acting profession stands out as the one that is most attacked. This may be related to the constant public attention and exposure that people in this industry experience, as well as the extreme scrutiny and criticism that their private lives and public appearances draw.

Singing Industry (12,245 comments): Singers, much like actors, frequently find themselves in the spotlight, with their performances, personal lives, and public statements often under scrutiny. The substantial amount of comments directed against singers highlights the difficulties artists face while navigating online environments, where harassment and criticism may be extremely harsh.

Politicians (4,590 comments): Given the divisive nature of politics and the public scrutiny that accompanies holding public office, internet abuse of politicians is nothing new. Politicians are frequently the target of criticism, insults, and even threats in the contentious political debate that occurs online, which is reflected in the comparatively large amount of comments directed against them.

Sports Personalities (2,653 comments): Athletes and sports figures also face a considerable amount of online scrutiny, especially in the age of social media where fans and critics alike can voice their opinions publicly. Comments targeting sports personalities may range from critiques of their performance to personal attacks based on factors such as race, gender, or behavior on and off the field.

Scholars (1,239 comments): Scholars and intellectuals may not receive as much media attention as celebrities or politicians, but they are not immune to internet abuse. When academics, researchers, or scholars participate in public debate or support contentious concepts, they may become the focus of internet trolls or those who have opposing opinions.

Other Professions (4,645 comments): The presence of comments targeting individuals from various other professions highlights the broad spectrum of online harassment, extending beyond just the realms of entertainment, politics, and sports. From healthcare professionals to educators, from business leaders to artists, individuals from diverse fields encounter online harassment, reflecting the widespread nature of this issue.

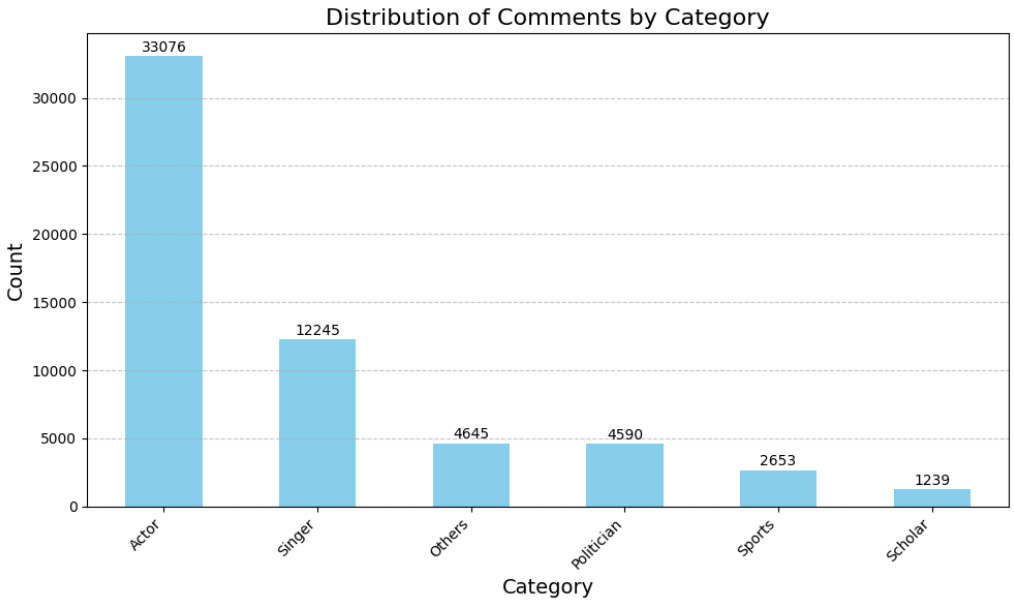


Figure 3.4: Category Distribution

Overall, the information shows the urgent need for comprehensive measures to combat online harassment in all professions, as well as the need of creating a safer and more inclusive online environment for all persons, regardless of vocation or professional background. Such remedies might include strong moderation systems on

social media sites, education and awareness efforts, and legal frameworks to hold online harassers responsible. By tackling online harassment holistically, we can help to create a digital world in which people may freely express themselves and engage in constructive discourse without fear of being harassed or threatened.

3.2.3 Label distribution

The distribution of labels, representing various types of bullying, within the dataset provides insight into the prevalence and nature of online harassment. Of the 58,448 comments evaluated, each labelled category reflects a separate facet of online behaviours, presenting a deeper picture of the issues provided by bullying and harassment online.

Sexual Comments: This form of online harassment is especially concerning since it frequently includes offensive language, unwanted approaches, and comments that disparage someone's gender or sexual orientation. Recipients of such remarks may have severe psychological impacts, including objectification, humiliation, and even trauma. Furthermore, sexual harassment on the internet has the potential to turn into physical threats or acts of violence, which is why it is a major worry for people's safety as well as for larger cultural standards. Strong moderation guidelines on digital platforms are necessary to address sexual remarks, but education and awareness programs are also necessary to combat detrimental attitudes towards gender and sexuality. Approximately 17.19% of the dataset, or 10,046 comments, are sexual in nature. These comments highlight a concerning aspect of online interactions that involve harassment or sexually explicit content. This disturbing trend in online communication is reflected in the comments, which frequently criticise people based on their gender or sexual orientation.

Not Bully Comments: While this category may appear benign at first look, it's important to remember that even seemingly neutral remarks can contribute to a hostile online environment. Even comments that are categorised as "not bullying" have the potential to reinforce discriminatory practices, stereotypes, and microaggressions that target certain people or groups. Furthermore, the sheer number of remarks that fit this description highlights how commonplace online interactions are that, even when they are not overtly aggressive, they may nevertheless be harmful to polite and inclusive conversation. Therefore, in order to address the underlying dynamics of online interactions, it is imperative that we promote a culture of empathy, critical thinking, and polite communication. In contrast, the label "not bully" includes the majority of the dataset's remarks, accounting for around 32.98% (19,269 comments). These comments are free of personal insults or harassment, reflecting a generally neutral or positive tone of debate.

Troll Comments: Deliberate provocation, mocking, or interruption with the intention of evoking strong emotions or sabotaging conversations is what defines trolling behaviour. Even while trolls don't always mean well when they attack, their acts can nonetheless cause psychological damage and create a hostile atmosphere for sincere communication. Trolls are difficult to deal with using traditional moderation

techniques because they frequently thrive on anonymity and the seeming impunity of online venues. A multifaceted strategy that includes community involvement, technical advancements, and platform regulations intended to deter and lessen disruptive conduct is needed to combat trolls. Troll comments, which account for around 25.15% of the dataset (14,702 comments), illustrate a common type of online abuse that consists of unpleasant contempt or mockery directed at individuals. These remarks help to create a toxic online atmosphere in which people may feel isolated or targeted by abusive behaviours.

Religious Comments: Online religious conversations may rapidly turn into heated disputes or downright hatred, especially when people express intolerant or radical beliefs. Religious remarks that promote hate speech, encourage violence, or degrade specific religious groups polarise online communities and exacerbate intergroup disputes. Combating religious harassment online demands a sophisticated strategy that values free expression while also supporting ideals of tolerance, mutual respect, and conversation. Platforms must find a balance between promoting varied viewpoints and preventing harmful content that breaches community norms or legal requirements. Religious comments, making up approximately 14.87% of the dataset with 8,689 comments, expose the intersection of online discourse with sensitive religious beliefs and identities. Often conveying offensive sentiments or spreading hatred towards religious groups, these comments underscore the need to address religious intolerance and promote respectful dialogue in digital spaces.

Threat Comments: Perhaps the most overt form of online harassment, threat comments pose a direct risk to individuals' safety and well-being. Whether they involve explicit threats of physical harm, intimidation, or coercion, such comments can instill fear and anxiety in their recipients, leading to real-world consequences such as stalking, doxxing, or offline violence. Effectively addressing threat comments requires swift and decisive action by platform moderators, including the removal of offending content, suspension or banning of perpetrators, and cooperation with law enforcement when necessary. Additionally, providing support and resources for victims of online threats is essential in mitigating the psychological toll of harassment and empowering individuals to seek help and protection. Approximately 9.83% of the dataset consists of 5,742 threat comments, which highlight the seriousness of online harassment by making explicit threats of violence or harm to specific persons. These comments highlight the possible real-world consequences of online misconduct and the significance of protecting persons from harm in digital contexts.

Finally, the distribution of labels within the dataset emphasises the complex and diverse character of online harassment, which includes a variety of abusive behaviours with far-reaching consequences for individual well-being, community dynamics, and social norms. Addressing these difficulties necessitates a multifaceted strategy that includes technology innovation, regulatory reforms, community participation, and education activities targeted at creating a safer, more inclusive digital environment for all users. Understanding the underlying mechanics of online harassment and applying proactive steps to avoid and mitigate harmful behaviour can help us create a digital world in which people can express themselves freely, politely, and without fear of harassment or intimidation. Overall, the distribution of labels showcases

the multifaceted nature of online bullying and harassment, encompassing various forms of abusive behavior. Addressing these concerns necessitates comprehensive measures for promoting online safety, encouraging polite communication, and countering harmful behaviours across digital platforms. To establish a safer and more inclusive online environment for all users, policies, community moderation, and education campaigns should be used in tandem.

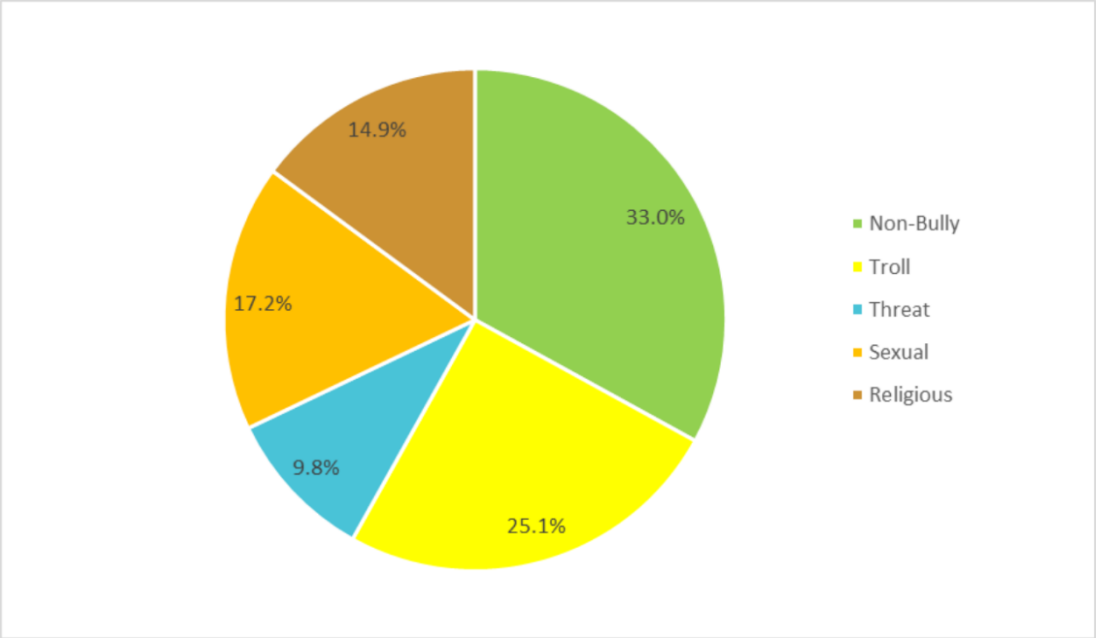


Figure 3.5: Label Distribution (Type of Bully)

Chapter 4

Methodology

4.1 Preprocessing

This section describes the steps taken to prepare data for analysis or model training. Preprocessing is an important part of natural language processing (NLP) tasks. It entails converting raw text data into a format that can be analyzed or used as input for machine learning models. The preprocessing procedures described here ensure that the data is clean, consistent, and suitable for a wide range of applications and models.

4.1.1 Dataset Cleaning

Various types of noise and irregularities are frequently present in text data, includingmissing values, emojis, non-Bengali characters and excessive whitespace. Text cleaning must be done thoroughly to remove these items since they can negatively impact machine learning model performance. Our method of text cleaning is based on the 'preprocess-bengali-text' function. It addresses several facets of text cleaning methodically. This guarantees that the final text data is devoid of superfluous symbols and appropriately formatted for additional processing. Emojis are transformed into textual representations, residual emoji textual representations are removed, non Bengali letters are filtered out and excess whitespace is removed in order to guarantee that the text data is standardized and free of any unnecessary components that can confuse the model.

4.1.2 Tokenization with BERT Tokenizer

One essential preprocessing step is tokenization. It divides textual data into smaller pieces known as tokens. These tokens function as the fundamental building blocks for deep learning model training, which helps the models comprehend and effectively interpret textual data. We tokenize every cleaned Bengali comment by using the BERT tokenizer. To guarantee consistent input size throughout all samples, this procedure includes segmenting the text into tokens. Then add special tokens to indicate the start and finish of each sequence. It also handles the task of managing padding and truncation. Furthermore, attention masks are produced to show which words are real and which ones are just padding tokens. Through the use of

this method, the model is better able to draw conclusions from the text data by concentrating its attention on pertinent tokens during training and inference.

4.1.3 Dataset Splitting

Separate datasets must be used for training, validation, and testing in order to conduct effective model evaluation. We may make sure that each step of the model generation process has its own collection of data by splitting the dataset. Three different subsets (i.e., training, validation and testing sets) were created from the preprocessed dataset. Where 80% of the dataset is selected for training, 10% is selected for validation and 10% is taken for testing. Consequently, the model's performance is assessed during training and hyperparameters are adjusted on the validation set. Lastly, the model's generalisation ability is assessed on untested data with the testing set. This partitioning technique guarantees an objective assessment of the model's performance. Additionally, this helps avoid overfitting and sheds light on the model's capacity to generalise to new, untested data.

4.1.4 Label Encoding and Class Weight Computation

For model compatibility, category labels (such as specific kinds of harassment) in text classification tasks must be transformed into numerical representations. Categorical labels are converted into numerical representations using label encoding, which enables the model to analyse and forecast using the encoded labels. We apply label encoding using the 'LabelEncoder' from scikit-learn. This provides a distinct numerical label to each category in the dataset. Other than that, class weights are calculated using the 'compute-class-weight' function to solve possible concerns with class imbalance. By preventing biases towards dominating classes while supporting equal proficiency across all categories, these class weights assist the model in suitably correcting for the unequal distribution of label classes during training.

4.2 Model Training Setup

This section outlines the configuration used to train our deep learning model for text classification in Bengali. Starting cross-validation, loading a pre-trained BERT model and setting up hardware resources for effective computing are all included in the procedure.

4.2.1 Cross-Validation Setup

Stratified k-fold cross-validation is utilised to guarantee a strong assessment of our model's performance and ability to generalise. The distribution of label classes inside each fold is preserved when the dataset is divided into k folds of equal size using this method. We have used the StratifiedKFold class from the scikit-learn library. We opted for five folds to ensure thorough evaluation, a common practice in many machine learning tasks. Each fold underwent training for a fixed number of epochs, with our choice being 20 epochs per fold. Through this approach, we aimed to obtain a comprehensive understanding of our model's performance across different subsets of the data. By assessing its consistency in performance across multiple

folds, we could gain insights into its generalization capabilities and potential areas for improvement.

4.2.2 Pre-Trained BERT Model Initialization

A BERT (Bidirectional Encoder Representations from Transformers) model that hasbeen trained and optimized for sequence classification tasks in order to classify texts in Bengali. The 'bert-base-multilingual-cased' version of BertForSequence-Classification from the transformers library is mainly used to initialise our model. This decision is driven by BERT's capacity to identify complex language patterns, whichis vital for our classification task.

4.2.3 GPU/CPU Configuration

We dynamically distribute hardware resources by detecting GPU availability to accelerate model training and take advantage of parallel processing capabilities. GPU calculation is utilized if one is available; otherwise, CPU execution is used by default. To provide smooth integration with both GPU and CPU architectures, the 'torch.device' module is employed in this function. To further enable effective use of computational resources, we relocate our initialized model to the designated device using '.to(device)'.

4.3 Model Training Initialization

The model training initialization procedure is described in this part, which includes setting up early stopping variables, optimization, configuring the scheduler, and allocating data structures to monitor training progress.

4.3.1 Early Stopping Initialization

We use early stopping based on validation loss to reduce the chance of overfitting and accelerate convergence. The following variables are initialised: 'best-val-loss': Initialised to (float('inf')), positive infinity, signifying the best observed validation loss throughout training. 'no-improvement': Started at zero and counts the number of epochs in which the validation loss has not improved. 'best-model-weights': A deep copy of the state dictionary of the original model that keeps the model's parameters that resulted in the lowest validation loss.

4.3.2 Optimization and Scheduler Setup

We use the AdamW optimizer for model optimisation, which is a Transformerspecific version of the Adam optimizer with weight decay regularisation. We set the learning rate (lr) to 2 × 105 and the epsilon () to 108, as it is generally advised for fine-tuning pre-trained BERT models. Besides, we designate 20 counts as the total number of training epochs.

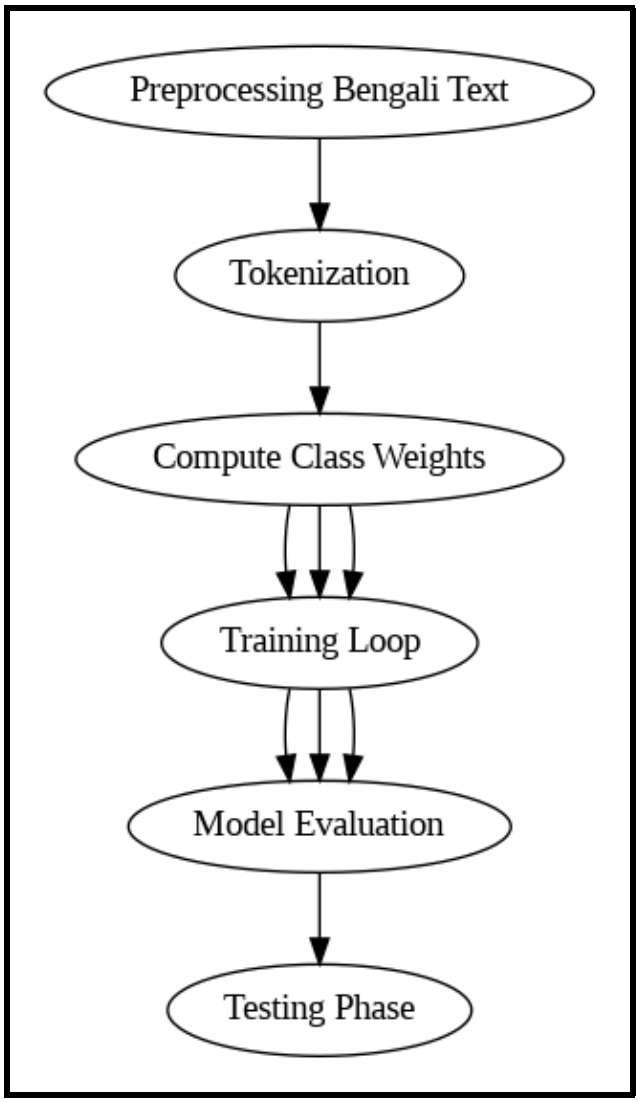


Figure 4.1: Workflow

4.4 Model Architecture

The model architecture used here is BERT (Bidirectional Encoder Representations from Transformers), specifically the "bert-base-multilingual-cased" variant[20]. BERT is a transformer-based model designed to understand the context of words in a sentence by looking at both preceding and following words, hence the term "bidirectional".

1. Input Representation

- **Token Embeddings:** BERT splits words into smaller subword units through WordPiece tokenization. Each subword within the text is then given a numerical value that represents its meaning and context.
- **Segment Embeddings:** BERT takes in pairs of sentences—like premise-hypothesis and question-answer pairs. BERT separates the two sentences in a pair using seg-

ment embeddings to give each sentence a special embedding.

- **Position Embeddings:** Unlike other sequence models, BERT does not inherently know the order of the words in a phrase. BERT understands the sequential nature of words by adding position embeddings to the embedding of each token.

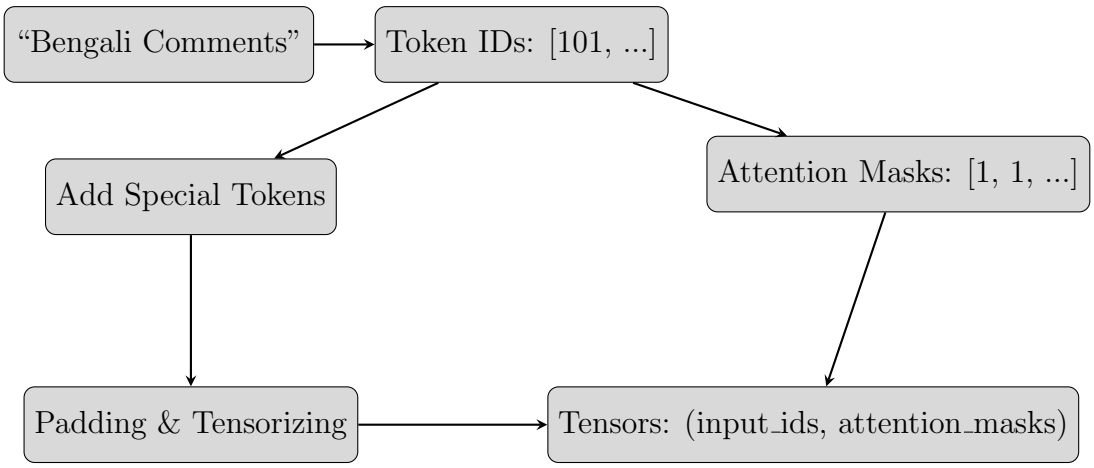


Figure 4.2: Input Representation

2. Transformer Layers

- **Multi-Head Self-Attention Mechanism:** Using the multi-head self-attention mechanism, BERT calculates the representation of each word by focusing on different segments of the input sequence. To effectively assess the importance of individual words inside the phrase and collect a variety of contextual elements, BERT uses several attention heads.

- **Feed-Forward Neural Network:** BERT sends the output to a feed-forward neural network once the self-attention mechanism has processed the input sequence. This network enhances the representations by nonlinearly changing the input.

- **Residual Connections and Layer Normalization:** BERT combines layer normalization and residual connections in every transformer block to enhance the training process and keep it stable. Through layer normalization, the inputs to each layer are standardized; residual connections allow gradients to flow during training, thus preventing the gradient problem.

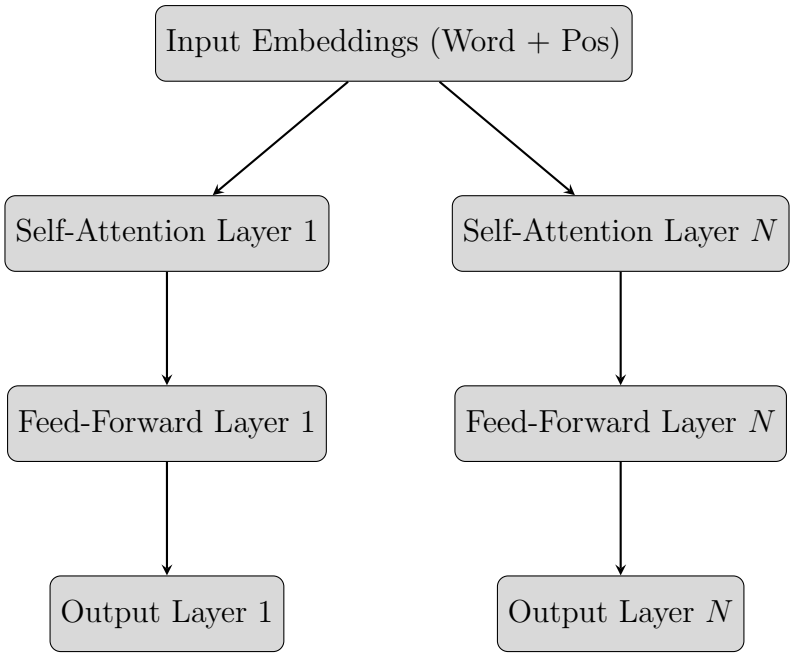


Figure 4.3: Transformer Layers

3. Output Representations

- **[CLS] Token:** BERT adds a special classification token at the beginning of the input sequence. The final hidden state of this token acts as an aggregate representation of the full input sequence. This representation becomes the foundation for making the prediction in tasks like question answering, sentiment analysis, or text classification.
- **[SEP] Token:** Used to separate sentences for tasks on pairs of sentences, such as text entailment or paraphrase identification. This helps BERT understand the structure of the input sequence and how to differentiate between sentences in a pair.

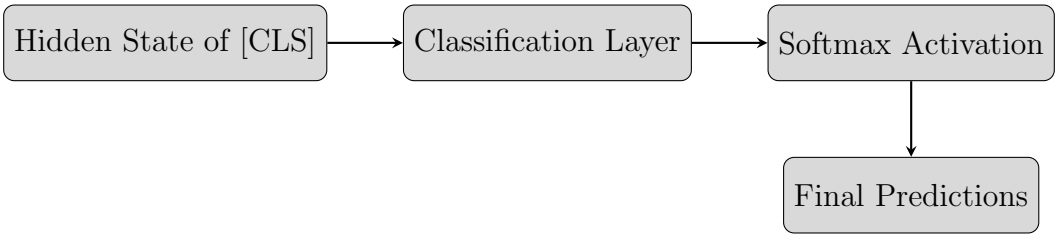


Figure 4.4: Output Representation

Sequence Classification with BERT

1. **Input Layer:** In sequence classification tasks, input sequences are first tokenized and converted into token, segment, and position embeddings.

2. **BERT Encoder:** After passing these embeddings through multiple layers of the transformer, BERT has the capacity to identify semantic links and contextual meaning in the input sequence.

3. **Classification Layer:** The final hidden state from the last transformer layer is passed through a fully connected classification layer. This layer produces the predicted probabilities for sequence classification by mapping the extracted representation to the output classes.

Chapter 5

Result Analysis

5.1 Training Phase

Classification reports offer a nuanced evaluation of the model's performance by providing insights into various metrics such as precision, recall, and F1-score for each class. By scrutinizing these metrics, we gain valuable insights into the model's ability to accurately classify instances across different categories.

5.1.1 Fold-wise Classification Reports

Learning about the performance of our model at different folds may aid in our comprehension of the model's dependability and effectiveness in Bengali text data classification. Let's now take a closer look at the classification reports for each fold:

Fold 1: With an accuracy of roughly 88.99%, Fold 1 has strong performance across several domains. Measures of recall, accuracy, and F1-score provide evidence for the model's ability to accurately recognize Bengali text data.

```
Fold 1 Accuracy: 0.8899700598802395
Fold 1 Classification Report:
              precision    recall  f1-score   support

   not bully       0.88      0.93      0.90      6160
   religious       0.92      0.91      0.92      2774
      sexual       0.88      0.91      0.89      3212
      threat       0.92      0.85      0.88      1850
       troll       0.89      0.82      0.85      4708

    accuracy                           0.89     18704
   macro avg       0.90      0.89      0.89     18704
weighted avg       0.89      0.89      0.89     18704
```

Figure 5.1: Fold 1

Not Bully: With a precision of 0.88 and a recall of 0.93, the model has a great abil-

ity to distinguish between non-bullying scenarios. The F1-score of 0.90 indicates a neutral performance in this area, with 6160 cases correctly classified as non-bullying. Thus, when the algorithm predicts a text to be "not bully," it makes the accurate prediction approximately 88% of the time and properly recognizes 93% of all texts that do not pose a threat to others. The precision-recall balance indicates that the model reduces false positives but may overlook some instances of language that is not bullying-related.

Religious: The precision, recall, and F1-score metrics are consistently good in this category, with values of 0.92, 0.91, and 0.92, respectively. The model's excellent accuracy in identifying 2774 instances of religious text classification demonstrates how precise it is. As a result, when the algorithm identifies a text as "religious," it accurately classifies texts in 91% of all religious scenarios and is roughly 92% correct. While the high precision indicates that the model hardly ever incorrectly identifies nonreligious materials as religious, the recall score indicates that the model properly recognizes most religious texts.

Sexual: Precision and recall ratings of 0.88 and 0.91, respectively, show that the model can identify sexual content. The F1-score of 0.89 indicates a balanced performance, with 3212 instances correctly classified as sexual text. This shows that the model captures 91% of all sexual encounters and accurately predicts sexual content 88% of the time. Based on the trade-off between recall and precision, the model seems to effectively identify sexual content while minimizing false positives.

Threat: Despite a somewhat lower recall of 0.85, the model maintains a high precision of 0.92 in identifying dangerous information. 1850 incidents are correctly classified as threatening with an F1-score of 0.88, demonstrating a great balance between recall and precision. This means that the model correctly detects 85% of all threatening occurrences and correctly predicts threatening content about 92% of the time. Even though the recall score indicates that there are some missing cases, the high precision ensures that there are few false positives in the classification of risky language.

Troll: Precision and recall scores of 0.89 and 0.82, respectively, show that the model can identify trolling content. The F1-score of 0.85, which correctly classified 4708 instances as trolling material, highlights a balanced performance. This demonstrates that 89% of the time, the algorithm correctly predicts trolling content and accounts for 82% of trolling instances. Accurate detection of trolling content is guaranteed by the high precision, even though there might have been some incidents that were missed.

The support, F1-score, and total accuracy statistics for Fold 1 are 18704, 0.89, and 0.89, respectively. These metrics provide a comprehensive assessment of the model's performance across a number of categories.

Fold 2: Fold 2 achieved an accuracy of around 96.46%, which is quite good. This astounding precision demonstrates the model's ability to accurately and consistently distinguish between various types of Bengali text input.

```
Fold 2 Accuracy: 0.9645511415280971
Fold 2 Classification Report:
              precision    recall  f1-score   support

   not bully       0.98      0.97      0.97      6160
   religious       0.96      0.98      0.97      2774
      sexual       0.96      0.96      0.96      3211
      threat       0.96      0.97      0.96      1850
       troll       0.95      0.95      0.95      4708

    accuracy                           0.96     18703
   macro avg       0.96      0.97      0.96     18703
weighted avg       0.96      0.96      0.96     18703
```

Figure 5.2: Fold 2

Not Bully: With remarkable precision, recall, and F1-score values of 0.98, 0.97, and 0.97, respectively, the model identified instances as not bullying. In other words, 98% of the time the model accurately labels a text as "not bully," and 97% of all instances that do not include bullying are included in the model's capture. While the high precision suggests a low percentage of false positives, the high recall demonstrates that the model can correctly identify the bulk of non-bullying occurrences.

Religious: Recall and precision metrics for recognizing religious text are 0.98 and 0.96, respectively, which is similarly outstanding. The model's accuracy in this category is further highlighted by the F1-score of 0.97, which shows that 2774 cases were accurately identified as religious. These measures show that the model performs exceptionally well at recognizing religious text while retaining a high degree of recall and precision.

Sexual: With an F1-score, recall, and precision of all 0.96, the model does well in the classification of sexual content. This suggests that the model correctly captures 96% of all sexual texts while maintaining a high degree of prediction precision. The balanced precision-recall trade-off shows how effective the model is at accurately and flawlessly detecting sexual material.

Threat: The precision and recall scores for identifying threatening material are both very good at 0.96, with an F1-score of 0.96. This demonstrates a balanced performance in appropriately identifying threatening scenarios, with 1850 examples correctly assessed as threatening. The model successfully detects the majority of dangerous messages and lowers the amount of false positives with a high degree of precision.

Troll: The precision and recall metrics for identifying trolling content are both very good at 0.95, with an F1-score of 0.95. After 4708 instances were effectively

identified, the model consistently classified instances as trolling text. These data demonstrate that the system correctly detects trolling content while maintaining a balanced precision-recall trade-off.

The support, F1-score, and overall accuracy scores for Fold 2 are 18703, 0.96, and 0.96, respectively. These numbers show how effectively the model works in detecting text data in Bengali across numerous categories with an impressive degree of accuracy.

Fold 3: Fold 3 outperformed in several categories, including precision, recall, and F1-score metrics, with an accuracy of almost 97.11%. The model's consistent and reliable performance shows that it can successfully classify Bengali text input.

```
Fold 3 Accuracy: 0.9710741592257927
Fold 3 Classification Report:
              precision    recall  f1-score   support

   not bully       0.97      0.99      0.98      6160
   religious       0.99      0.97      0.98      2774
      sexual       0.97      0.97      0.97      3212
      threat       0.98      0.96      0.97      1850
       troll       0.97      0.95      0.96      4707

    accuracy                           0.97     18703
   macro avg       0.97      0.97      0.97     18703
weighted avg       0.97      0.97      0.97     18703
```

Figure 5.3: Fold 3

Not Bully: Excellent precision, recall, and F1-score are obtained when classifying instances as non-bullying, with values of 0.97, 0.99, and 0.98, respectively. Consequently, when the algorithm correctly predicts a text to be "not bully," it does so about 97% of the time, successfully identifying 99% of all communications that do not involve bullying. The model's ability to correctly identify the vast majority of cases that are not bullying is demonstrated by its high recall, and its high precision suggests that there are not many false positives.

Religious: The precision and recall metrics for identifying religious information are notable as well, scoring 0.99 and 0.97, respectively. With an F1-score of 0.98, the model correctly identified 2774 examples as religious, underscoring its accuracy in this domain. The results of these tests demonstrate that the model has remarkable memory and precision when it comes to identifying religious text.

Sexual: This model performs well in categorizing sexual content, with an F1-score, recall, and precision of 0.97. This suggests that while accurately classifying 97% of all sexual texts, the model retains a high level of prediction accuracy. Precision-recall trade-off is balanced, indicating that the model is effective in accurately and flawlessly detecting sexual material.

Threat: The precision and recall scores for identifying threatening material are excellent, both at 0.98, with an F1-score of 0.97. The capacity to recognize threatening circumstances is demonstrated by the balanced performance in properly recognizing 1850 occurrences as threatening. By correctly recognizing the majority of dangerous messages while maintaining a high level of precision, the model reduces false positives.

Troll: These results are outstanding, with an F1-score of 0.96 and precision and recall metrics for identifying trolling content of 0.97 and 0.95, respectively. The model consistently performed well in this domain, successfully identifying 4707 instances of trolling material. These data demonstrate that the system correctly detects trolling content while maintaining a balanced precision-recall trade-off.

The values that correspond to Fold 3 are- 18703 for support, 0.97 for F1-score, and 0.97 for total accuracy. The aforementioned metrics illustrate the model's effectiveness in consistently classifying Bengali textual data into many groups.

Fold 4: With an exceptional accuracy of about 97.49% in Fold 4, the model demonstrated excellent precision, recall, and F1-score metrics across several categories. The model's high degree of accuracy and robustness in classification tasks is highlighted by this level of performance, which makes it a dependable tool for Bengali text analysis.

```
Fold 4 Accuracy: 0.9748703416564187
Fold 4 Classification Report:
              precision    recall  f1-score   support

   not bully       0.99      0.98      0.98      6160
   religious       0.98      0.98      0.98      2774
      sexual       0.96      0.98      0.97      3212
      threat       0.95      0.98      0.97      1850
       troll       0.98      0.96      0.97      4707

    accuracy                           0.97     18703
   macro avg       0.97      0.98      0.97     18703
weighted avg       0.98      0.97      0.97     18703
```

Figure 5.4: Fold 4

Not Bully: With scores of 0.99, 0.98, and 0.98, respectively, the precision, recall, and F1-score for classifying incidents as non-bullying are excellent. It can be seen from this that the model accurately predicts a text as "not bully" around 99% of the time and accounts for 98% of all cases that are not bullying. The high recall shows the model's capacity to accurately identify the great majority of non-bullying cases, while the high precision suggests few false positives.

Religious: With an F1-score of 0.98, the precision and recall metrics for recognizing religious material are equally remarkable, measuring 0.98 for both. This highlights

how accurate the model is in this category, properly identifying 2774 occurrences as religious. These measures show that the model performs exceptionally well at recognizing religious text while retaining a high degree of recall and precision.

Sexual: With an F1-score of 0.97, recall of 0.98, and precision of 0.96, the model performs well in identifying sexual material. This implies that the model maintains a high degree of prediction precision while successfully capturing 96% of all sexual texts. The model's efficacy in precisely and impeccably identifying sexual material is demonstrated by the balanced precision-recall trade-off.

Threat: The precision and recall scores for identifying threatening material are both outstanding, at 0.95 and 0.98, respectively, with an F1-score of 0.97. The capacity to recognize threatening circumstances is demonstrated by the balanced performance in properly recognizing 1850 occurrences as threatening. By correctly recognizing the majority of dangerous messages while maintaining a high level of precision, the model reduces false positives.

Troll: These results are outstanding, with an F1-score of 0.97 and precision and recall metrics for identifying trolling content of 0.98 and 0.96, respectively. The model consistently performed well in this domain, successfully identifying 4707 instances of trolling material. These data demonstrate that the system correctly detects trolling content while maintaining a balanced precision-recall trade-off.

The support, F1-score, and total accuracy results for Fold 4 are 18703, 0.97, and 0.97, respectively. These metrics show how effectively the model works in classifying text data in Bengali across multiple categories with consistency.

Fold 5: With an astounding precision of approximately 98.37%, Fold 5 demonstrated remarkable performance. The model's effectiveness in successfully classifying Bengali text instances was highlighted by this fold, which displayed high precision, recall, and F1-score metrics across many categories.

```
Fold 5 Accuracy: 0.9837459231139389
Fold 5 Classification Report:
              precision    recall  f1-score   support

    not bully       0.99      0.99      0.99      6160
    religious       0.98      0.99      0.98      2773
       sexual       0.98      0.99      0.98      3212
       threat       0.98      0.99      0.98      1850
        troll       0.98      0.98      0.98      4708

     accuracy                           0.98     18703
    macro avg       0.98      0.98      0.98     18703
 weighted avg       0.98      0.98      0.98     18703
```

Figure 5.5: Fold 5

Not Bully: With scores of 0.99, 0.99, and 0.99, respectively, the precision, recall, and F1-score for classifying incidents as non-bullying are remarkable. This shows that the algorithm accurately predicts a text as "not bully" approximately 99% of the time and accounts for 99% of all cases that aren't bullying. The high recall shows that the model can accurately identify the great majority of non-bullying events, while the high precision means that there aren't many false positives.

Religious: Both the precision and recall metrics for identifying religious material are remarkably high, measuring 0.98 for both, with an F1-score. The fact that the algorithm accurately classified 2773 instances as religious shows how accurate it is in this area. Based on these metrics, the model demonstrates remarkable performance in identifying religious material while maintaining a high recall and precision level.

Sexual: The model does a good job at identifying sexual content, with an F1-score of 0.98, recall of 0.99, and precision of 0.98. This suggests that while accurately identifying 98% of all sexual texts, the model keeps a high level of prediction precision. The model balances precision and memory, demonstrating its effectiveness in accurately and flawlessly classifying sexual material.

Threat: The precision and recall scores for identifying threatening material are excellent at 0.98 and 0.99, respectively, with an F1-score of 0.98. The capacity to recognize threatening circumstances is demonstrated by the balanced performance in properly recognizing 1850 occurrences as threatening. By correctly recognizing the majority of dangerous messages while maintaining a high level of precision, the model reduces false positives.

Troll: These results are outstanding, with an F1-score of 0.98 and precision and recall metrics for identifying trolling content of 0.98 and 0.98, respectively. The model consistently performed well in this domain, correctly identifying 4708 instances of trolling language. These data demonstrate that the system correctly detects trolling content while maintaining a balanced precision-recall trade-off.

The support, F1-score, and total accuracy results for Fold 5 are 18703, 0.98, and 0.98, respectively. These metrics show how effectively the model works in classifying text data in Bengali across multiple categories with consistency. Finally, Fold 5 demonstrates how well the model retains high accuracy and balanced performance metrics despite being able to distinguish between different Bengali text categories. This level of performance reveals the model's reliability and effectiveness in real-world scenarios, making it a handy tool for many NLP jobs.

5.2 Evaluation of Trained Bengali Text Categorization Model

In this section, we detailed the evaluation of our trained Bengali text categorization model. During the testing phase, we conducted a thorough analysis of several distinct elements, such as classification reports, accuracy metrics, and a comprehen-

sive examination of the confusion matrix. Our aim was to obtain a comprehensive understanding of the model's performance in accurately identifying Bengali text data.

5.2.1 Accuracy Assessment

Analyzing the performance of our model effectively was crucial to determining its dependability and usefulness. We learned a lot about its advantages and shortcomings by examining its categories and evaluating a variety of measures.

5.2.1.1 Testing Accuracy: A Measure of Reliability

The core of our evaluation was the testing accuracy of our model, which came in at approximately 93.47%. This metric served as a compass, enabling us to see how frequently the vast majority of Bengali text data that was accessible was correctly predicted by our model. Stated differently, our model achieved an outstanding level of precision, correctly identifying approximately 93 or 94 out of every 100 samples it encountered.
However, accuracy might not have given us a clear view of our model's performance on its own. More research was needed to understand its classification across various Bengali text categories.

5.2.1.2 Class-wise Evaluation: Precision, Recall, and F1 Score

A closer look at our data provided information on how well the model performed in each of the following text categories:

- **Not Bully**: With a precision, recall, and F1 score of 94%, our model was quite good at recognizing texts that were not bullying. This showed that approximately 94% of the time, our algorithm was accurate when it stated that a text did not involve bullying. Furthermore, it effectively caught 94% of all non-bullying incidents, reducing the possibility that we may have missed these communications. Our model had a good grasp of this area, correctly identifying 3919 examples of non-bullying language.

- **Religious**: Precision, recall, and F1 score for our model's identification of religious text were all 94%, indicating consistency. This indicated a high degree of accuracy in identifying religious texts. Our algorithm demonstrated its proficiency in navigating the subtleties of this category by accurately detecting 1715 occurrences of religious literature.

- **Sexual**: Our model accurately and precisely recognized approximately 93% of sexual content, with a recall and precision of 94%. This showed that it accurately identified 94% of all sexual communications while decreasing false positives. The car's 94% F1 score indicated that it did well in this category. Our model demonstrated its capacity to handle sensitive data by properly identifying instances of sexual text messages from 2015.

- **Threat**: One of our model's best features was its 94% accuracy rate in detecting potentially dangerous data. By striking a balance between accuracy and

recall, it successfully gathered 93% of all threatening texts. After accurately recognizing 1086 instances of frightening content, our algorithm showed that it had a good understanding of this category.

- **Troll**: Our model demonstrated remarkable accuracy in identifying trolling content, with precision, recall, and F1 score of over 92%. This implied that it correctly detected 92% of all texts containing trolling while minimizing false positives. With an accuracy rate of 2955 for trolling text, our model showed that it could navigate the complexities of this category.

5.2.1.3 Evaluation and Interpretation

By combining results from independent category evaluations, we fully understood the operation of our model. With an average precision, recall, and F1 score of over 94%, our model had a high ability to accurately detect Bengali text across multiple categories.

Based on 11690 examples examined, this comprehensive review emphasized the practicality and efficacy of the concept. By striking a balance between recall and precision across several text categories, our algorithm effectively recognized nuances and classified Bengali text data.

Our model's great comprehension of Bengali text was demonstrated by its balanced performance across categories and high accuracy, which made it a useful tool for tasks like topic categorization, sentiment analysis, and content moderation. However, in order to adjust to changing user behaviors and linguistic patterns, ongoing monitoring and updates were required.

```
Testing Accuracy: 0.9347305389221557
Testing Classification Report:
              precision    recall  f1-score   support

   not bully       0.94      0.94      0.94      3919
   religious       0.94      0.94      0.94      1715
      sexual       0.93      0.94      0.94      2015
      threat       0.94      0.93      0.94      1086
       troll       0.92      0.92      0.92      2955

    accuracy                           0.93     11690
   macro avg       0.94      0.94      0.94     11690
weighted avg       0.93      0.93      0.93     11690
```

Figure 5.6: Testing Phase Result

Looking back at our accuracy evaluation, we could see that our model had potential for Bengali text classification. We needed to keep improving its algorithms, adding more training data, and incorporating user feedback to further maximize its potential.

5.3 Overall Metrics

The model exhibits an overall accuracy of 93.47%, showcasing its capacity to classify comments across different categories accurately. This high level of accuracy is instrumental in supporting content moderation efforts and ensuring a safer online environment by effectively distinguishing between varying types of online behavior.

In this section, we describe in detail the evaluation of our trained Bengali text categorization model. To do this, we perform a thorough analysis of several distinct elements, such as classification reports, accuracy metrics, and a thorough look at the confusion matrix. With this systematic investigation, we want to obtain a comprehensive understanding of the model's performance in accurately identifying Bengali text data.

First, we look at accuracy metrics to assess the overall performance of the model. These measurements provide useful information on the percentage of correctly identified occurrences when compared to the total number of examples studied. By examining accuracy scores, we may ascertain the model's capacity to provide accurate predictions across different classes in the Bengali text dataset.

Furthermore, we obtain detailed information on the F1-score, precision, and recall for each class using classification reports. Exactness elucidates the model's ability to recognize examples of a specific class from all occurrences accurately classified as that class. On the other hand, recall indicates how well a model can identify each and every instance of a class among all of the examples that belong to that class. The F1-score strikes a balance between recall and precision to give a fair evaluation of a model's performance for a particular class.

Furthermore, we do a comprehensive analysis of the confusion matrix, which offers a graphical depiction of the classification performance of the model in various classes. We can find regions where the model may need to be further refined and patterns of misclassifications by closely examining the confusion matrix. We can identify particular classes that present difficulties for the model and come up with ways to improve its precision and resilience thanks to this thorough examination.

Essentially, we can rigorously and precisely evaluate the effectiveness of our trained Bengali text classification model thanks to our extensive evaluation system. We obtain important insights through the use of accuracy metrics, classification reports, and the confusion matrix. These insights guide iterative enhancements and optimizations, which eventually improve the model's capacity to correctly classify Bengali text input.

5.4 Confusion Matrix Analysis

A model's confusion matrix is a crucial instrument for assessing the degree to which it can classify data. The graphical depiction of the model's performance allows us

to view the distribution of accurate and inaccurate classifications across different classes, providing us with a more complete picture of its performance. By closely scrutinizing this matrix, we analyze the subtleties of the model's decision-making process and uncover patterns and trends that illustrate both the benefits and drawbacks of the model. When we iteratively refine and optimize the model, these realizations serve as benchmarks that help us push the boundaries of precision and dependability further by fine-tuning its parameters and practices. In a sense, the confusion matrix serves as a compass, pointing us in the direction of continued improvement and ensuring that the model is still able to manage the challenges posed by real-world data.

When we look at the confusion matrix, we can see that most cases are correctly identified, indicating strong overall performance. However, misclassification rates are higher in some classes than in others; this is particularly evident when comparing labels like "sexual," "religious," and "not bully." This discrepancy highlights potential areas for improvement and emphasizes how important it is to improve the model's accuracy and robustness.

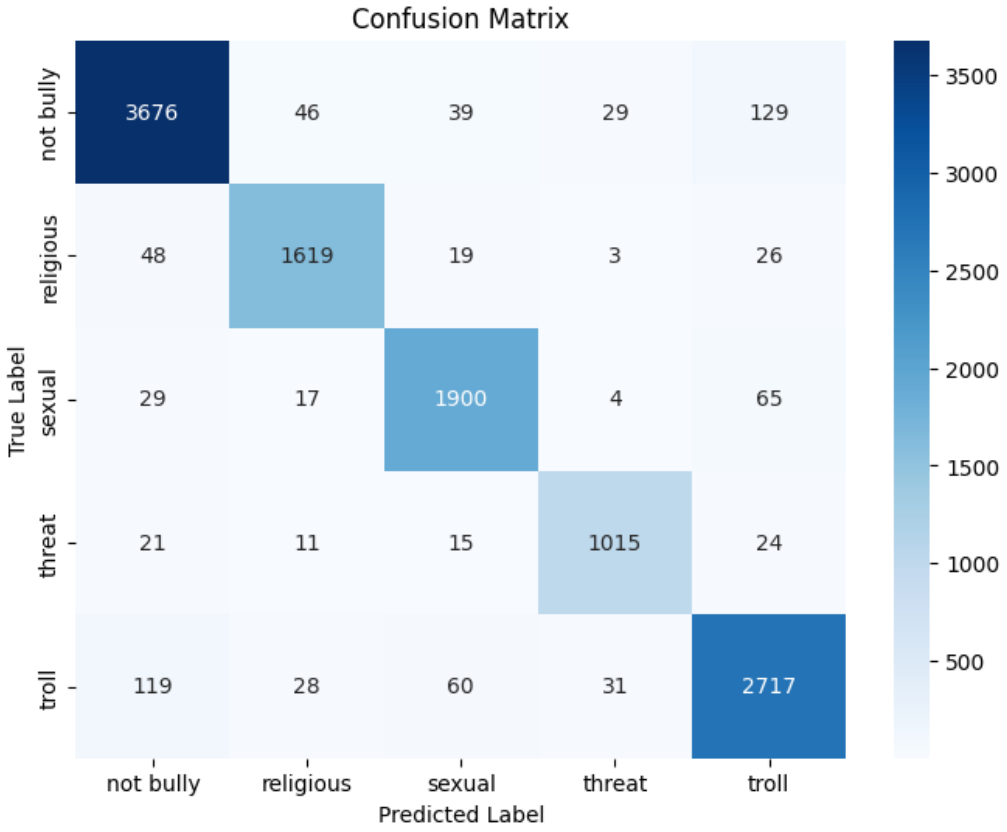


Figure 5.7: Confusion Matrix

The confusion matrix has rows that represent the actual class and columns that indicate the anticipated class. For this reason, diagonal elements show proper classifications, while off-diagonal elements show misclassifications. By examining these misclassifications in great detail and recognizing patterns and tendencies, we are able to pinpoint specific areas that require improvement.

Such misclassified categories as "not bully," "religious," and "sexual" suggest that the model has to improve in its ability to discern even the smallest characteristics within these categories. Incorporating additional features or enhancing the model architecture may be necessary to more accurately convey the unique characteristics of every class. Furthermore, we may discover areas that require further work by examining the confusion matrix, which provides information about the model's performance in other categories. We can iteratively improve the model's accuracy and effectiveness by taking a methodical approach to misclassification patterns.

In summary, analyzing the confusion matrix sheds light on the model's classification performance and enables us to enhance and fortify its accuracy and robustness. By leveraging these insights and implementing targeted enhancements, we can optimize the model's utility and impact and ensure that it remains dependable and successful in real-world applications.

5.5 Training and Validation Losses

The plot shows the trend of training and validation losses over successive epochs. It provides insights into how well the model is learning from the training data and whether it is overfitting or underfitting. Ideally, both training and validation losses should decrease steadily over epochs. A large gap between the training and validation losses may indicate overfitting, where the model performs well on the training data but fails to generalize to unseen data.

Figure 5.8: Training and Validation Losses

5.6 Training and Validation Accuracies

The plot displays the training and validation accuracies across epochs. It depicts how accurately the model predicts the labels for both the training and validation

datasets. Similar to the loss plot, the accuracy plot should show increasing trends for both training and validation datasets. However, if the training accuracy significantly surpasses the validation accuracy, it could indicate overfitting.

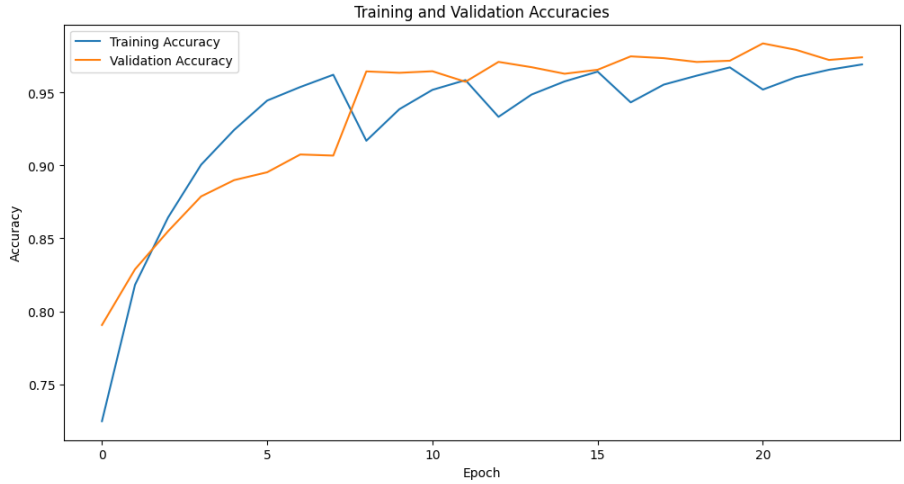


Figure 5.9: Training and Validation Accuracies

By analyzing these plots together, we can gain insights into the model's learning dynamics, identify potential issues such as overfitting or underfitting, and make informed decisions about model optimization and tuning parameters. Overall, the aim is to achieve low training and validation losses along with high training and validation accuracies, indicating a well-generalized model capable of accurately predicting labels for unseen data.

5.7 Comprehensive Performance Analysis of Online Behavior Classification Model

The model's capability to classify online comments into specific behavior categories—Non Bully, Religious, Sexual, Threat, and Troll—is evaluated using precision, recall, and F1-score metrics. Each metric provides insights into the model's accuracy and reliability in identifying relevant content, crucial for effective content moderation and maintaining a safe online environment.

Non Bully Class: Precision at 94.43% indicates a high accuracy level, with the model correctly identifying non-bullying comments. The recall of 93.80% suggests the model's effectiveness in capturing most of the true non-bullying instances, reducing false negatives. The F1-score of 94.12% demonstrates a balanced performance, indicating robustness in distinguishing non-bullying content.

Religious Class: Achieving a precision of 94.07%, the model accurately identifies comments related to religion. With a recall of 94.40%, it captures the majority of genuine religious comments, ensuring minimal oversight. The F1-score at 94.24% highlights a harmonious balance between precision and recall, reflecting the model's

reliability in classifying religious content.

Sexual Class: The precision score of 93.46% underscores the model's accuracy in pinpointing comments with sexual content. The recall rate of 94.29% shows its ability to detect a vast majority of actual sexual remarks, enhancing content monitoring. An F1-score of 93.87% indicates the model's effectiveness in balancing accuracy and coverage.

Threat Class: With a precision of 93.81%, the model demonstrates its proficiency in identifying comments containing threats. A recall of 93.46% ensures that it recognizes most real threatening remarks, critical for safety measures. The F1-score of 93.63% reflects a well-balanced detection capability, asserting the model's robustness in threat identification.

Troll Class: Precision stands at 91.76%, indicating the model's effectiveness in recognizing trolling behavior, despite being slightly lower than other classes. The recall score of 92.09% portrays the model's adeptness at detecting actual trolling comments, aiming to reduce false negatives. The F1-score of 91.88% signifies a balanced approach between precision and recall, confirming consistent performance.

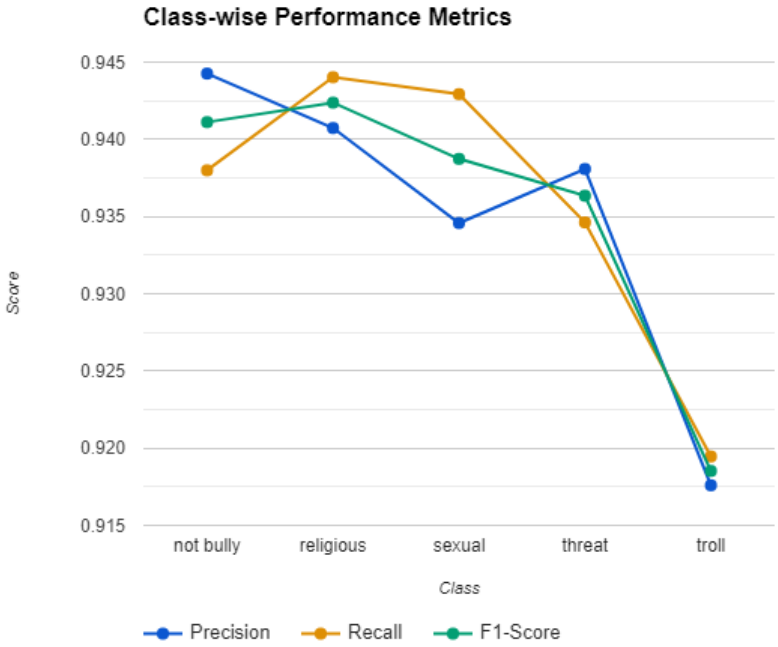


Figure 5.10: Class-wise Performance Metrics

5.8 Compute ROC curve and ROC area for each class

The evaluation of the model's performance is a crucial step in understanding its ability to identify hazardous information in online comments. By means of an extensive analysis, we want to provide an understanding of the general performance of the model in several categories, as well as its benefits and drawbacks.

The model's ability to discriminate between different types of problematic remarks needs to be assessed carefully. By thoroughly analyzing the system's performance metrics, such as precision, recall, and the Area Under the Curve (AUC), we may evaluate how well it performs in accurately classifying comments that are thought to be potentially dangerous. Important details regarding the discriminative ability of the model are provided by Receiver Operating Characteristic (ROC) curves and other visual aids. The True Positive Rate (TPR) and False Positive Rate (FPR) trade-off is depicted in these curves, providing a comprehensive view of the system's performance at different thresholds. These ROC curves produce AUC values, which are numerical representations of the model's efficacy. Higher AUC values, which show how well the model can correctly classify comments into the relevant categories, are indicative of more discriminative ability. High AUC values were seen in every class, according to our research, which shows excellent overall performance. The capacity of the model to correctly categorize comments into their appropriate categories is highlighted by the ROC curves, which show high TPRs while keeping low FPRs.

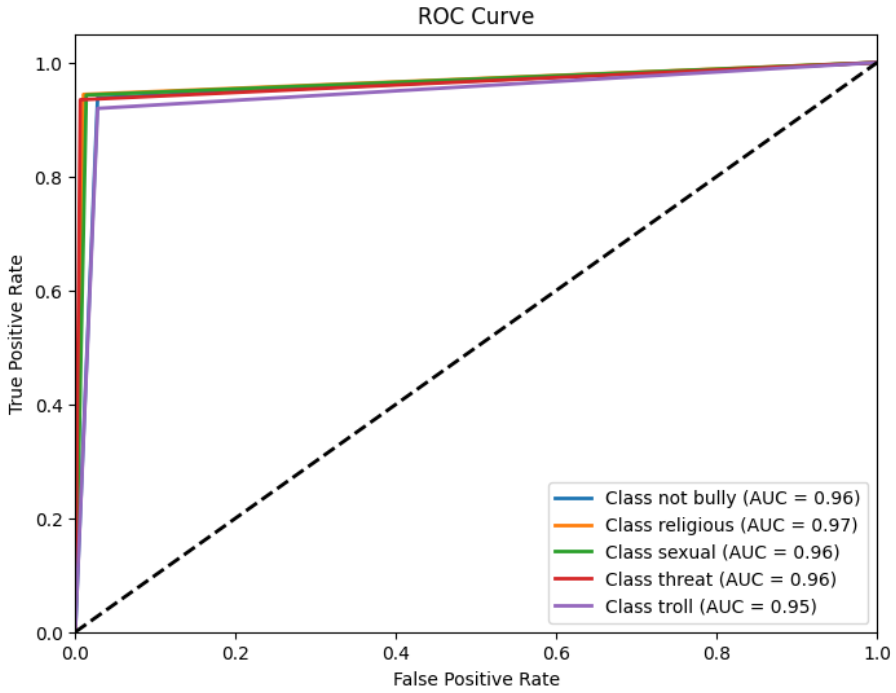


Figure 5.11: ROC curve

By efficiently differentiating between various categories of problematic content in

online comments, the model exhibits good discriminative performance, with AUC values ranging from 0.95 to 0.97. These results support the model's effectiveness in spotting and highlighting potentially dangerous remarks, which helps to create a safer online community.

5.9 Comparison between Classes

A comparison of classes can provide important insights into how well the model performs with specific kinds of problematic content. By assessing the degree to which it is able to identify various types of harmful comments, we can gain additional insights about its pros and cons.

Notably, comments about religious material had the highest AUC values. "Not Bully," "Sexual," and "Threat" are the next most closely followed categories. The AUC value for remarks associated with bullying is marginally lower, despite the model's solid performance in accurately identifying such remarks. While the frequency of comments may vary between classes, the model consistently performs well across all categories. Its ability to reliably and adaptably identify hazardous information in real-world contexts is demonstrated by its resilience to class imbalance and ability to effectively identify information regardless of class distribution.

5.10 Insight into Class Imbalance

The phenomenon known as class imbalance, defined as an unbalanced distribution of data among multiple classes, greatly hinders machine learning model training and evaluation. Predictive bias and inaccurate model performance may result from some classes in the dataset being overrepresented in comparison to others. All things considered, our thorough examination shows that the model manages this difficulty admirably, exhibiting strong and reliable performance in every class.

The model performs consistently and reliably even though the frequency of comments from each class in the dataset fluctuates. The model can successfully distinguish between positive and negative samples within each class, as seen by the high Area Under the Curve (AUC) values found in all categories. Regardless of class distribution, the model's strong AUC values show that it can reliably identify problematic content in a variety of categories. The model displays objectivity and effectiveness in reliably detecting and classifying harmful information inside online comments, as seen by consistently high AUC values across all classes.

In addition, the model's robustness against class imbalance is highlighted by its capacity to maintain consistent performance across all classes. Unlike methods that could be biased in favor of the majority class, our analysis shows that the model remains neutral and objective in its predictions, ensuring reliable detection of problematic content in all categories. This strong performance is crucial in real-world applications where the distribution of comments across different categories may vary significantly. By addressing the issues brought about by class imbalance,

the methodology enhances its applicability and dependability in a range of settings, facilitating the more successful detection and mitigation of online toxicity.

5.11 Area under AUC

When evaluating the effectiveness of machine learning models, the Area Under the Curve (AUC) is a crucial metric to consider, particularly for binary classification tasks like finding harmful content in online comments. It offers a thorough assessment of the model's capacity for discrimination and predictability across a range of thresholds. The model has performed exceptionally well, according to our research, with AUC values ranging from 0.95 to 0.97. These values show that the model can discriminate between events that are positive, or problematic, and those that are negative, or benign, for each class. These high AUC values demonstrate the model's good discriminatory capacity by demonstrating that it accurately scores positive instances higher than negative ones across a wide range of decision criteria.

These results carry significant implications, particularly in the context of online content management. Dangerous content must be recognized and removed in the modern digital age, since online platforms are the primary means of expression and communication. By accurately distinguishing between harmful and innocuous comments, the approach empowers platform management to address issues such as hate speech, harassment, and cyberbullying before they arise, so fostering a safer and more inviting online environment for users. The model's consistent performance in every class is additional evidence of its dependability and efficacy. The model retains its discriminatory power to guarantee accurate identification of problematic content in a range of circumstances, even in cases where there may be differences in the frequency of comments in different categories. This resistance to class imbalance is especially interesting since it shows that the model's ability to learn from and generalize from the data, rather than the distribution of instances within the dataset, determines the model's success.Furthermore, the impressive AUC values obtained in each class show how flexible and adaptive the model is. While certain machine learning algorithms may do better than others at identifying specific types of harmful information, such as hate speech or explicit language, others may not. However, our methodology is effective in a wide range of contexts, including sexual content, bullying, threats, and intolerance for religious beliefs, indicating that it can be used to a wide range of online toxicity.

In summary, the model's outstanding performance—which is demonstrated by high AUC values in every class—confirms its value as a useful instrument for content moderation efforts. Platform administrators are able to prevent harmful conduct and keep ahead of developing problems by using the model, which leverages rigorous evaluation metrics like AUC and powerful machine learning algorithms. By doing this, it helps to advance the development of an online community that is safer, more welcoming, and more positive for people all around the world.

Chapter 6

Conclusion

With a focus on the Bengali language, the research work presented tackles the crucial issue of identifying toxic language and cyberbullying on social media platforms. The goal of the study is to use a reliable algorithm that can recognize instances of cyberbullying in social media posts written in Bengali. The specific objective is to overcome the special difficulties presented by the Bengali language and the paucity of available resources by developing a framework that is linguistically and culturally appropriate. The dataset used in this study consists of 58,448 Bengali comments from a variety of social media sites, such as Facebook, Instagram, and YouTube. These remarks are divided into five groups: threat, religious, troll, non-bully, and sexual, to achieve the research objective.

The transformer model 'bert-base-multilingual-cased' was employed. Our results show an accuracy higher than 95% on average across different folds for the classification of Bengali text data. Furthermore, the precision, recall, and F1 scores of all categories are above 90%. Moreover, by analyzing the fold-wise classification reports, it was found that each fold had indicators of accurate and strong performances. This insight provides that the model is consistent and reliable across all sets of data. The major contributions and results from this project prove the extreme limit of understanding or achievement in the categorization of Bengali text data, mainly in the field of content moderation and analysis of online activities. Our work provides a milestone in the research and development of Bengali language cyberbullying detection. The study highlights the efficacy of transformer models in detecting cyberbullying in Bengali while also pointing out the challenges in expanding and adapting to new and evolving forms of cyberbullying.

The system will be adjusted to function with new internet platforms and communication channels, and it will be made more scalable to manage larger data volumes in the future. In addition, the study aims to continuously improve the dataset and work on different algorithms to ensure their effectiveness in preventing new types of cyberbullying. The ultimate goal is to improve detection technologies and raise awareness about the consequences of cyberbullying to make the internet a safer environment for Bengali-speaking users.

Bibliography

[1] S. Kemp, *Digital 2023: Bangladesh - datareportal – global digital insights*, Feb. 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-bangladesh.

[2] N. Shahbazi, Y. Lin, A. Asudeh, and H. V. Jagadish, "Representation bias in data: A survey on identification and resolution techniques," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, Jul. 2023, ISSN: 1557-7341. DOI: 10.1145/3588433. [Online]. Available: http://dx.doi.org/10.1145/3588433.

[3] X. Ying, *An overview of overfitting and its solutions*, Feb. 2019. DOI: 10.1088/1742-6596/1168/2/022022.

[4] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 1096–1100. DOI: 10.1109/ICESC48915.2020.9155700.

[5] M. Behzadi, I. G. Harris, and A. Derakhshan, "Rapid cyber-bullying detection method using compact bert models," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 2021, pp. 199–202. DOI: 10.1109/ICSC50631.2021.00042.

[6] M. Gada, K. Damania, and S. Sankhe, "Cyberbullying detection using lstm-cnn architecture and its applications," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–6. DOI: 10.1109/ICCCI50826.2021.9402412.

[7] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, *Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques*, Nov. 2021. DOI: 10.3390/electronics10222810. [Online]. Available: http://dx.doi.org/10.3390/electronics10222810.

[8] B. Haidar, C. Maroun, and A. Serhrouchni, *A multilingual system for cyberbullying detection: Arabic content detection using machine learning*, Dec. 2017. DOI: 10.25046/aj020634.

[9] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, *Social media cyberbullying detection using machine learning*, 2019. DOI: 10.14569/IJACSA.2019.0100587. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100587.

[10] D. Chatzakou, I. Leontiadis, J. Blackburn, *et al.*, *Detecting cyberbullying and cyberaggression in social media*, 2019.

[11] A. Akhter, K. Uzzal, and M. Polash, *Cyber bullying detection and classification using multinomial naive bayes and fuzzy logic*, 2019.

[12] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. B. Ashraf, *Cyberbullying detection using deep neural network from social media comments in bangla language*, 2021. arXiv: 2106.04506 [cs.CL].

[13] M. G. Hussain, T. A. Mahmud, and W. Akthar, "An approach to detect abusive bangla text," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, 2018, pp. 1–5. DOI: 10.1109/CIET.2018.8660863.

[14] K. R. Talpur, S. S. Yuhaniz, and N. Amir, *Cyberbullying detection: Current trends and future directions*, 2020.

[15] S. Sarker and A. R. Shahid, *Cyberbullying of high school students in bangladesh: An exploratory study*, 2018. arXiv: 1901.00755 [cs.CY].

[16] Z. Alsaed and D. Eleyan, *Approaches to cyberbullying detection on social networks: A survey*, Jul. 2021.

[17] R. Ghosh, S. Nowal, and G. Manju, *Social media cyberbullying detection using machine learning in bengali language*, 2021.

[18] M. I. H. Emon, K. N. Iqbal, M. H. K. Mehedi, M. J. A. Mahbub, and A. A. Rasel, "Detection of bangla hate comments and cyberbullying in social media using nlp and transformer models," in *Advances in Computing and Data Sciences: 6th International Conference, ICACDS 2022, Kurnool, India, April 22–23, 2022, Revised Selected Papers, Part I*, Springer, 2022, pp. 86–96.

[19] P. G. N. Neogi, A. H. F. Fahim, F. K. Khan, F. K. K. Khan, and M. F. F. Faisal, *Dataset: Bengali online comments*, Zenodo, May 2024. DOI: 10.5281/zenodo.11295622. [Online]. Available: https://doi.org/10.5281/zenodo.11295622.

[20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[21] M. A. Moreno, A. D. Gower, H. Brittain, and T. Vaillancourt, *Applying natural language processing to evaluate news media coverage of bullying and cyberbullying*, 2019.

[22] R. Kumar, B. Lahiri, and A. K. Ojha, *Aggressive and offensive language identification in hindi, bangla, and english: A comparative study*, 2021.

[23] M. Das, S. Banerjee, P. Saha, and A. Mukherjee, *Hate speech and offensive language detection in bengali*, 2022. arXiv: 2210.03479 [cs.CL].

[24] M. T. Ahmed, M. Rahman, S. Nur, A. Islam, and D. Das, "Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021, pp. 1–10. DOI: 10.1109/ICAECT49130.2021.9392608.

[25] Abdhullah-Al-Mamun and S. Akhter, "Social media bullying detection using machine learning on bangla text," in *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, 2018, pp. 385–388. DOI: 10.1109/ICECE.2018.8636797.

[26] S. Sultana, M. O. F. Redoy, J. Al Nahian, A. K. M. Masum, and S. Abujar, *Detection of abusive bengali comments for mixed social media data using machine learning*, 2023.