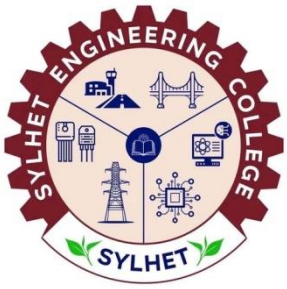
SYLHET ENGINEERING COLLEGE

Affiliated with Shahjalal University of Science and Technology

Department of Computer Science & Engineering



Bangla Hate Speech Detection: A Multi-Platform Machine Learning Approach

Submitted by

Taahia Tahsin
Reg. No.:2019331521

Musammat Tania Sultana Hafsa
Reg. No.:2019331532

Department of Computer Science & Engineering

Supervisor

Md. Abu Naser Mojumder
Assistant Professor
Head of the CSE Department
Sylhet Engineering College, Sylhet

Recommendation Letter from Thesis Supervisor

The thesis entitled “Bangla Hate Speech Detection: A Multi Platform Machine Learning Approach” submitted by the students

1. **Taahia Tahsin**
2. **Musammat Tania Sultana Hafsa**

is a record of research work carried out under my supervision and I, hereby, approve that the report is submitted in partial fulfilment of the requirement for the award of their Bachelor’s Degree.

Signature of the Supervisor
Md. Abu Naser Mojumder
Assistant Professor
Head of the CSE Department
Sylhet Engineering College, Sylhet

Acknowledgements

First and foremost, we offer our deepest gratitude to the Almighty, whose boundless mercy and silent guidance have been our constant source of strength throughout this journey. Through every challenge and every success, His blessings have illuminated our path and sustained our resolve.

We would like to express our heartfelt appreciation to our respected supervisor, **Md. Abu Naser Mojumder**, for his invaluable support, continuous encouragement, and insightful feedback. His guidance has been pivotal in shaping the direction and quality of our research, and we are truly grateful for his mentorship.

Our sincere thanks also go to our respected teachers, **Md. Lysuzzaman** and **Nayan Kumar Nath**, whose dedication to teaching and depth of knowledge have left a lasting impact on our academic foundation. Their support and encouragement throughout our studies have been a source of inspiration.

We are especially thankful for the strong collaboration and mutual understanding we shared as thesis partners. The joint effort in completing our thesis, " **Bangla Hate Speech Detection: A Multi-Platform Machine Learning Approach**" was made possible through shared dedication, teamwork, and a commitment to learning.

Lastly, with all our love and gratitude, we acknowledge the endless support of our families. Their unconditional love, sacrifices, and constant prayers have been the backbone of our academic journey. Without their unwavering belief in us, this achievement would not have been possible.

Bangla Hate Speech Detection: A Multi Platform Machine Learning Approach

by

Taahia Tahsin, Musammat Tania Sultana Hafsa

Submitted to the Department of Computer Science & Engineering, in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science & Engineering

Abstract

The growing amount of hate speech on social media platforms emphasizes the urgent need for strong detection systems are needed, especially in low-resource languages like Bangla where contextual cues like irony, sarcasm, humor, and threat present difficult classification problems. By creating a manually curated dataset of 10,000 Bangla posts and comments gathered from Facebook, Instagram, and Twitter, this thesis addresses three major research limitations: platform-specific focus, inadequate contextual modeling, and the absence of extensive annotated Bangla datasets. Using binary hate/non-hate labels and multi-phase manual validation, each entry was meticulously annotated. It was then divided into fine-grained groups such as humor, sarcasm, irony, and menace to capture both overt and covert types of hate speech.The dataset reflects real-world usage with diverse linguistic forms including Romanized Bangla, code-mixed content, and informal expressions. Six different models—SVM, BiLSTM, CNN+BiLSTM, BanglaBERT, mBERT, and XLM-RoBERTa—were trained and evaluated using stratified sampling and comprehensive preprocessing. Among them, BanglaBERT achieved the highest performance (87% accuracy, 0.86.7 F1-score), benefiting from monolingual language-specific pretraining. mBERT (85.5%) and XLM-RoBERTa (85%) also performed competitively with strong multilingual capabilities. SVM outperformed several deep models with a notable 77% accuracy, highlighting its efficiency in linear separability with handcrafted features. BiLSTM and CNN+BiLSTM yielded 74% and 72% accuracy respectively, struggling particularly with complex linguistic patterns. Detailed error analysis revealed that transformer-based models consistently outperformed others in detecting implicit hate forms like sarcasm and irony, although humor-based hate speech remains a challenging category across all models. This thesis not only contributes one of the largest annotated Bangla hate speech corpora to date but also provides practical insights into model selection for low-resource languages, demonstrating that transformer-based models—especially language-specific ones—are highly effective for context-aware hate speech detection in multilingual and informal settings. The outcomes of this research provide a solid foundation for future Bangla NLP work and reinforce the importance of ethically-driven and context-sensitive modeling in combating online hate.

Contents

Chapter 1

Introduction

1.1 Background

As user-generated content on social media platforms has grown in popularity in the digital age, hate speech has spread widely, endangering societal harmony, public safety, mental health, and individual dignity. Social media sites like Facebook, Twitter, and Instagram have become into havens for bigotry, threats, and disparaging comments. In addition to targeting people or groups because of their identity, affiliations, or views, hate speech also creates division, feeds negative stereotypes, and in severe situations, incites actual violence. The need for strong, automated hate speech detection systems has become critical due to its quick dissemination, which has been fostered by anonymity, virality, and ineffective moderation.

While machine learning (ML) and natural language processing (NLP) have shown promise in automating hate speech detection, most existing systems are tailored to high-resource languages like English and depend heavily on superficial cues such as keyword matching or sentiment polarity. These methods often fail to capture the deeper, more complex aspects of hate speech that involve cultural context, implicit meaning, or indirect expressions such as sarcasm, coded language, humor, and irony. The challenge becomes more pronounced in under-resourced languages like Bangla, where linguistic diversity, informal language, Romanized script, and code-mixing further complicate detection.

Moreover, the lack of publicly available, large-scale, and well-annotated datasets in Bangla has been a significant barrier to progress in this field. Without high-quality training data, even state-of-the-art models underperform, particularly when faced with implicit or context-dependent expressions. The constantly evolving nature of online language, regional dialects, and culturally rooted expressions further limits the effectiveness of static or monolingual approaches. It is also important to consider the multilingual nature of online discourse in Bangladesh, where users often switch between Bangla, English, and Romanized Bangla within a single post. As such, any effective detection system must be linguistically flexible and capable of understanding mixed-language input.

To address these challenges, this thesis adopts a multi-platform, linguistically informed strategy for Bangla hate speech detection. A significant contribution of this research is the creation of a manually curated dataset comprising 10,000 Bangla comments and posts collected from Facebook, Twitter, and Instagram. These entries capture a wide range of hate and non-hate content, including nuanced categories such as sarcasm, irony, humor, and threat, enabling the development of models that can better understand both explicit and implicit hate expressions. This work not only fills a critical resource gap for Bangla NLP

but also lays the foundation for developing inclusive and context-aware hate speech detection systems that are scalable and applicable to real-world moderation challenges.

1.2 Problem Statement

Hate speech on social media platforms has become a major concern, particularly in low-resource languages like Bangla, where the lack of comprehensive datasets, contextual ambiguity, and informal linguistic styles make detection extremely challenging. Existing hate speech detection models are often built on English-language datasets and rely on platform-specific or keyword-based techniques that fail to generalize across Bangla's diverse digital expressions. These models also struggle to identify implicit hate forms such as sarcasm, humor, and irony, which are commonly used in Bangla social media discourse. Additionally, most available datasets are either too small, poorly annotated, or biased toward specific topics or sources. This leads to poor model performance, misclassification, and limited real-world applicability. Therefore, there is a pressing need for a context-aware, multilingual, and scalable hate speech detection system tailored to Bangla that can work across platforms and detect both explicit and implicit hate speech with higher accuracy.

1.3 Motivation

The increasing volume of hate speech on social media platforms has become a major concern in today's digitally connected world, especially in regions like Bangladesh where online discourse often reflects deep-rooted social tensions. Despite growing awareness, the tools available to detect and moderate hate speech in the Bangla language remain extremely limited. Most existing detection systems are either designed for high-resource languages like English or are incapable of identifying subtle and implicit forms of hate such as sarcasm, irony, and coded humor—forms commonly found in Bangla social media interactions.

The absence of large-scale, annotated datasets in Bangla further hampers progress in this area. Without high-quality, context-rich data, machine learning models struggle to perform effectively. Moreover, the linguistic diversity of online Bangla content—including the use of Romanized Bangla, code-mixing, and informal speech—poses additional challenges that are not adequately addressed by current solutions. These gaps in both resources and methodologies inspired the development of this thesis.

By manually curating a diverse and representative dataset of 10,000 Bangla comments from Facebook, Twitter, and Instagram and evaluating multiple deep learning models—including BiLSTM, BanglaBERT, XLM-RoBERTa, CNN-BiLSTM, mBERT, and SVM—this research aims to bridge these gaps. The goal is not only to improve the accuracy of hate speech detection in Bangla but also to build a scalable and ethical framework that can adapt to the evolving nature of online communication in low-resource languages. This work is driven by the need to contribute meaningful tools to combat digital hate and promote safer,

more inclusive online spaces for Bangla-speaking users.

1.4 Research Area

This thesis lies at the intersection of **Natural Language Processing (NLP)**, **Machine Learning (ML)**, and **Computational Social Science**, with a specific focus on **automated hate speech detection in low-resource languages**. It explores how advanced deep learning models, particularly transformer-based architectures, can be applied to detect hate speech in Bangla—a linguistically rich but underrepresented language in NLP research.

Key subfields involved in this research include:

- **Text Classification:** Binary and fine-grained classification of social media text into hate and non-hate categories, along with nuanced classes like sarcasm, irony, humor, and threat.
- **Multilingual and Code-Mixed NLP:** Addressing the challenges posed by Romanized Bangla, code-switching between Bangla and English, and informal, noisy user-generated content.
- **Social Media Analysis:** Mining and interpreting content from platforms like Facebook, Instagram, and Twitter, which are dynamic and culturally diverse sources of communication.
- **Low-Resource Language Processing:** Developing solutions for languages with limited annotated corpora and pretraining resources, using transfer learning and language-specific models.
- **Ethical AI and Online Safety:** Promoting digital well-being by building systems that can help moderate harmful content, reduce toxicity, and create safer online environments.

This research contributes to the growing body of work in context-aware NLP and demonstrates how data-driven, ethically grounded methodologies can address real-world problems in digital communication.

1.5 Research Aim

The primary aim of this research is to develop an accurate and context-aware hate speech detection and classification system for the Bangla language, using a manually curated multi-platform dataset and advanced deep learning models. The study seeks to effectively identify both explicit and implicit forms of hate speech—such as threats, sarcasm, irony, and humor—from social media platforms like Facebook, Twitter, and Instagram, and provide practical solutions for improving hate speech moderation in low-resource language settings.

1.6 Research Objectives

This thesis aims to develop an effective Bangla hate speech detection and classification system by addressing the limitations of existing approaches. The specific objectives of the research are as follows:

- Build a large annotated Bangla hate speech dataset (10K samples from Facebook, Twitter, Instagram).
- Label data with:
  - Hate/non-hate tags
  - Subtypes (sarcasm, irony, humor, threats)
- Train & compare models:
  - ML (SVM)
  - DL (BiLSTM, CNN+BiLSTM)
  - Transformers (BanglaBERT, mBERT, XLM-R)
- Evaluate performance using precision, recall, F1-score.
- Improve detection of sarcasm/irony using contextual models.

Chapter 2

Literature Review

2.1 Identified Gaps

The detection of hate speech has been extensively studied in recent years, with researchers employing various machine learning and deep learning techniques. Previous studies have primarily focused on single platforms, such as Twitter or Facebook, and have used datasets annotated for specific types of hate speech (e.g., racial, religious, or gender-based). For instance,

Mullah et al. (2021) [1] in their study *"Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review,"* provided a comprehensive analysis of machine learning techniques for hate speech detection on social media platforms. They categorize methods into **classical machine learning**, **ensemble learning**, and **deep learning** approaches. **SVM, Naïve Bayes, and Logistic Regression** are commonly used classical models, while **Random Forest and Boosting** enhance performance through ensemble techniques. Deep learning models, including **CNN, LSTM, and BiLSTM**, outperform traditional methods by capturing **contextual and implicit hate speech**. The study highlights **dataset challenges**, such as **limited size, bias, and lack of linguistic diversity**, affecting model generalization. Multi-platform datasets from **Twitter, Facebook, and other sources** are suggested to improve detection accuracy. **Deep learning models excel in understanding linguistic nuances**, but **SVM remains a strong baseline** for text classification. The study identifies **data imbalance and cultural variations** as major obstacles to accurate hate speech detection. It emphasizes the need for **better datasets, hybrid models, and bias mitigation strategies**. Overall, the review serves as a foundation for future research aiming to develop **robust and adaptable** hate speech detection systems.

Abro et al. [2] conducted a comparative study on automatic hate speech detection using machine learning techniques. The study evaluated **three feature engineering methods** (Bigram with TF-IDF, Word2Vec, Doc2Vec) and **eight classifiers** (SVM, NB, RF, KNN, DT, LR, MLP, AdaBoost) on a **CrowdFlower dataset of 14,509 tweets**. The **bigram with TF-IDF approach outperformed others**, and **SVM achieved the highest accuracy (79%)**, followed by AdaBoost and RF. KNN and MLP performed the worst due to inefficiency with noisy data. The research also highlighted **dataset class imbalance**, making hate speech harder to classify. The findings serve as a **baseline for future studies**, emphasizing the need for **larger, more balanced datasets and improved classification techniques**.

Omran et al. (2023) [3] compare various machine learning algorithms for hate speech

detection on social media, using techniques like N-grams, TF-IDF, and Bag-of-Words for feature extraction. The results show that the **Naïve Bayes + Decision Tree** model achieves the best balance of accuracy (**88.74%**) and efficiency, making it ideal for real-time applications. However, the study is **limited to Twitter data**, which may not generalize to other social media platforms with different linguistic patterns. Additionally, **deep learning models like BERT were not explored in depth**, and **biases in the dataset were not addressed**, which could impact fairness and model performance.

The paper "Hate Speech Classification Using SVM and Naïve Bayes" by Asogwa et al. [4] examines machine learning models for hate speech detection. It compares SVM and Naïve Bayes, finding that SVM achieves 99.37% accuracy, while Naïve Bayes performs poorly with 50%. The dataset, sourced from the UCI Machine Learning Repository, includes 62,485 instances of offensive and non-offensive speech. The study suggests exploring deep learning models and expanding the dataset to enhance model generalizability. Limitations include the exclusion of deep learning models and the use of a single dataset, limiting broader applicability.

Ayo et al. (2020) [5] reviewed various **machine learning techniques** for hate speech classification on Twitter, proposing a **metadata-based architecture** that achieved an **F1-score of 91.5%**. However, the study lacks **real-time validation** and does not address **threshold tuning and fragmentation issues**.

Khanday et al. (2022) [6] focused on **COVID-19-related hate speech detection**, using **traditional ML and ensemble learning**, where **Stochastic Gradient Boosting achieved 98.04% accuracy**. The study is **limited to COVID-19 data, lacks deep learning models, and does not evaluate real-time performance**.

Geet d'Sa et al. (2020) [7] used deep learning models like CNN, Bi-LSTM, and CRNN with fastText and BERT embeddings to classify hate speech. Their Twitter-based study showed that **fine-tuning BERT improved F1-score by 16% over traditional methods.** However, the model **struggles with detecting implicit hate speech and requires significant computational power.**

William et al. (2022) [8] compared different machine learning models for hate speech detection and found that SVM with bigram features performed best, achieving 79% accuracy. The study, based on the **CrowdFlower dataset,** highlights challenges in **detecting sarcasm and the absence of deep learning techniques.**

Jahan & Oussalah (2023) [9] reviewed various machine learning and deep learning approaches for hate speech detection. They found that **deep learning generally outperforms traditional models but still faces difficulties with multilingual and context-dependent hate speech.**

Sanoussi et al. (2022) [10] analyzed hate speech in Chadian French-Arabic Facebook comments and found that SVM with FastText embeddings achieved 95.4% accuracy. However, their study is **limited to Chadian text and does not include deep learning methods.**

Despite progress, most studies focus on single platforms, missing the full scope of hate speech

across different media. Existing methods also struggle with context and linguistic diversity. This thesis tackles these gaps by **building a diverse dataset from Facebook, Instagram, Twitter, and newspapers** and developing better feature engineering for more accurate hate speech detection in Bengali

Chapter 3

Methodology

This section describes the comprehensive methodological framework employed to detect and classify context-aware hate speech in Bangla using deep learning and machine learning models. Our approach integrates classical techniques like SVM, deep learning architectures (BiLSTM, CNN-BiLSTM), and transformer-based models (Bangla-BERT, mBERT, XLM-RoBERTa), aiming to handle nuanced contexts such as irony, sarcasm, humor, and threats present in Bangla social media content.

3.1 Dataset Description

The dataset used in this research was compiled from multiple Bangla sources, including newspapers, social media, and online comment sections. It contains over **10,081 Bangla textual instances**, distributed across different sheets in an Excel file. Each instance was labeled with:

- **Irony**: Hate expressed subtly through contradiction or misleading positivity.
- **Sarcasm**: Hurtful statements disguised as praise or humor.
- **Humor**: Offensive content hidden within jokes or satire.
- **Threat**: Direct intimidation or incitement of violence.

These categories were especially important in building a **context-aware classification system**, as many forms of hate speech in Bangla are indirect and culturally embedded.

3.1.1 Dataset Visualization

	A	B	C
1	Facebook		
2	**Sentence**	**Label**	**Category**
3	এই সরকার তো আবার উন্নয়নের জোয়ার এনেছে, রাস্তা খোঁড়া ছাড়া কিছুই তো হয় না!	0	Sarcasm
4	তুমি তো সত্যিই একজন বুদ্ধিমান, তোমার মতামত শুনে তো দেশের ভবিষ্যৎ উজ্জ্বল!	0	Irony
5	তোমার মত লোকদের জন্যই দেশটা আজ এই অবস্থায়, সাবধান হও!	0	Threat
6	"আজকে আবার বিদ্যুৎ চলে গেছে, মনে হয় সরকার আমাদের অন্ধকারে রাখতে চায়!"	0	Sarcasm
7	"তোমার পোস্ট দেখে হাসতে হাসতে পেট ব্যথা হয়ে গেল!"	1	Humor
8	"তুমি যদি আরেকটা মিথ্যা বলো, তাহলে কিন্তু ফল ভালো হবে না!"	0	Threat
9	"এই রকম বুদ্ধিমান মানুষদের জন্যই তো দেশ এগিয়ে যাচ্ছে!"	0	Irony
10	"তোমার মতামত শুনে মনে হচ্ছে তুমি সত্যিই একজন বিশেষজ্ঞ!"	0	Sarcasm
11	"তোমার ছবি দেখে মনে হচ্ছে তুমি মডেল হতে পারো!"	1	Humor
12	"তুমি যদি আবার এমন কিছু করো, তাহলে কিন্তু পুলিশে জানাবো!"	0	Threat
13	ওহ, তুমি তো বিশাল জ্ঞানী দেখছি, ফেসবুকে না এসে বই পড়লে ভালো হতো!	0	Sarcasm
14	আরে ভাই, তোমার কথায় দেশের উন্নতি হবে না, চা-দোকানে গল্প করে লাভ কি?	1	Humor
15	তুমি আবারও এমন মন্তব্য করলে, তোর খবর আছে!	0	Threat
16	সত্যি বলছি, এমন মেধাবী মানুষ দেখা যায় না আজকাল!	0	Irony
17	তোর টাইমলাইনে এমন পোস্ট দিয়ে কি লাভ, কেউ তো সিরিয়াসলি নেবে না!	1	Humor
18	তুমি কি চাঁদের বুকে বাড়ি বানাতে চাও? তোর স্বপ্ন দেখার ক্ষমতা তো অসীম!	1	Humor
19	তোর মত বেকুবদের জন্যই দেশে গরিবি বাড়ছে!	0	Hate
20	তোর কমেন্ট দেখে তো মনে হচ্ছে, তুই সরকারি চাকরি পেয়েই গেছিস!	0	Irony
21	মজা করে বলছি, তোর এই হাস্যকর পোস্ট একবারে মাস্টারপিস!	1	Humor
22	আরও একটা বাজে কথা বললে, কিন্তু তোর খবর নেবো!	0	Threat
23	এই যে হিরো সাহেব, একটু দয়া করে বাস্তবতায় ফিরে আস!	0	Sarcasm
24	তোর এই ধরনের কথা শুনে তো মানুষ হাসতে হাসতে মরে যাচ্ছে!	1	Humor
25	তোর পোস্ট পড়ে তো মনে হচ্ছে, তুই বিজ্ঞানী নাকি জ্যোতিষী?	0	Sarcasm
26	একটু চিন্তা করে লিখলে ভালো হতো, এমন আবোল-তাবোল পোস্ট না দিয়ে!	0	Irony
27	আহা, তুই তো দারুণ কমেডিয়ান হয়ে যাচ্ছিস দেখছি!	1	Humor
28	এমন বাজে পোস্ট আবার দিলে, তোর আইডি রিপোর্ট করবো!	0	Threat
29	তোর মত বেকুবদের জন্যই ফেসবুকের মান কমে যাচ্ছে!	0	Hate

Fig 2: Sample data collected from Facebook

	A	B	C
1	Instagram		
2	Sentence	Label	Category
3	এই পোস্টে যদি কেউ হাহা রিয়্যাক্ট দেয়, পুলিশে দেব!	0	Threat
4	আজকাল কার মেয়েরা শুধু রিল বানাতেই ব্যস্ত, পড়াশোনা তো পুরাই গেছেগা!	0	Humor
5	তুই তোর মতামত রাখ, আমরা তোকে সিরিয়াসলি নিই না!	1	Sarcasm
6	এই ধরনের হাস্যকর বক্তব্যে দেশ এগোবে কীভাবে?	0	Sarcasm
7	ভাই তোর কথায় এত জ্ঞান, তুই কি জীবনে কিছু করছিস	0	Irony
8	আজকাল যেকোনো পোস্টে সবাই বিশ্লেষক হয়ে যায়!	0	Sarcasm
9	তোর কমেন্ট দেখে মনে হচ্ছে তোকে হাসাতে আসছি আমরা	1	Humor
10	তুই যদি এতটাই জানিস, তাহলে নিজেই করে দেখাস না কেন?	0	Irony
11	এই পোস্টটা দেখে মনে হচ্ছে বাংলাদেশের ভবিষ্যৎ কাঁদছে!	0	Sarcasm
12	এই জাতির মানুষগুলো সমাজের জন্য বিষের মতো।	0	Threat
13	এই পোস্টে হাসি থামাতে পারছি না ভাই	1	Humor
14	তুই তো একেবারে শান্তির দূত, সবাইকে যুদ্ধ লাগিয়ে রাখিস	0	Irony
15	ওহ হো! তুমি তো একেবারে নীতির প্রতীক	0	Sarcasm
16	তোর মতামতের ভাণ্ডার খুলেই বসে থাকি	0	Irony
17	ওদের ধরে ধরে দেশছাড়া করো, জায়গা নেই এদেশে	0	Threat
18	তুই লিখিস, আমরা শিখি	0	Sarcasm
19	তোর পোস্ট মানেই বিনোদন গ্যারান্টি	1	Humor
20	তুই কমেন্ট করলেই পোস্টের মান বেড়ে যায়	0	Irony
21	আর একবার এইরকম বললে রিপোর্ট করব	0	Threat
22	তুই তো ভাইরাল হবার লোভে জাতির বিশ্লেষক	0	Sarcasm
23	তোর মত ইনফ্লুয়েন্সার পেয়ে ধন্য আমরা	1	Irony
24	এই পোস্টের নিচে তোর মতামত, জাতি আজ ধন্য	0	Sarcasm
25	তোর জ্ঞান দেখে গুগলও লজ্জায় পড়ে যাবে	0	Humor
26	মতামতটা এতই দরকারি, বই বের করা উচিত	0	Irony
27	তোর মতামতেই তো নীতি নির্ধারণ হয়	0	Sarcasm
28	কমেন্ট পড়ে মনে হচ্ছে সবাই কমেডিয়ান	1	Humor
29	তোর মতামত ছাড়া তো আমরা অন্ধ ছিলাম	0	Irony

Fig 2: Sample data collected from Instagram

	A	B	C
1	Tweeter		
2	Sentence	Label	Category
3	ওরে বেয়াদব, তোকে থামাতে সময় লাগবে না	0	Threat
4	আজকের টুইটটা অসাধারণ ছিল	1	Humor
5	তোর মত গাধা মানুষ টুইটারে এল কিভাবে?	0	Sacrcasm
6	এসব ফালতু মতামত দিয়া কারে প্রভাবিত করিস?	0	Irony
7	এমন সত্যি কথা বলার সাহস কম জনেরই থাকে	1	Irony
8	ভাইরে, আপনার টুইট দেখে আজ মন ভালো হয়ে গেল	1	Humor
9	তুই একেবারে দেশের বোঝা	0	Humor
10	তোর পোস্ট মানেই জাতির সর্বনাশ	0	Sarcasm
11	এত সুন্দরভাবে কথা বলা যায়?	1	Irony
12	আপনি পোস্ট দিলে জানি কিছু শিখবো	1	Humor
13	আপনার কথাগুলো বেশ কাঁটাছেড়া করলেও সত্যি	1	Sarcasm
14	তোর মতো আবালদের কারণে দেশের ক্ষতি	0	Humor
15	আজকের মতামতটা একদম অন পয়েন্ট	1	Irony
16	. এত সুন্দর ব্যাখ্যা আগে দেখিনি	1	Humor
17	ভাইয়ের ব্যঙ্গাত্মক টুইটেও গভীর সত্য	1	Sarcasm
18	এই মানুষটা রীতিমত কমেডি কিং	1	Humor
19	যদি তোকে একদিন পাই, ঠিক করে দিবো	0	Threat
20	রাজনীতির পেছনের কৌতুকও আছে	1	Humor
21	কথা শুনে হাসি পায়, কিন্তু বাস্তব চিন্তা জাগে	1	Irony
22	আপনার ট্রলগুলো সমাজের আয়না	1	Sarcasm
23	এমন বুদ্ধিদীপ্ত ট্রল রোজ চাই	1	Humor
24	আপনার লেখার স্টাইল একেবারে অন্যরকম	1	Irony
25	গাধা টাইপ লোকজন এখন পলিটিক্স শেখায়!	0	Irony
26	আপনার টুইট দেখে একসাথে হাসি আর ভাবনা আসে	1	Humor
27	এমন পোস্টগুলো মানুষকে ভাবায়	1	Irony
28	আপনার প্রতিটা টুইট মানেই নতুন চিন্তা	1	Humor

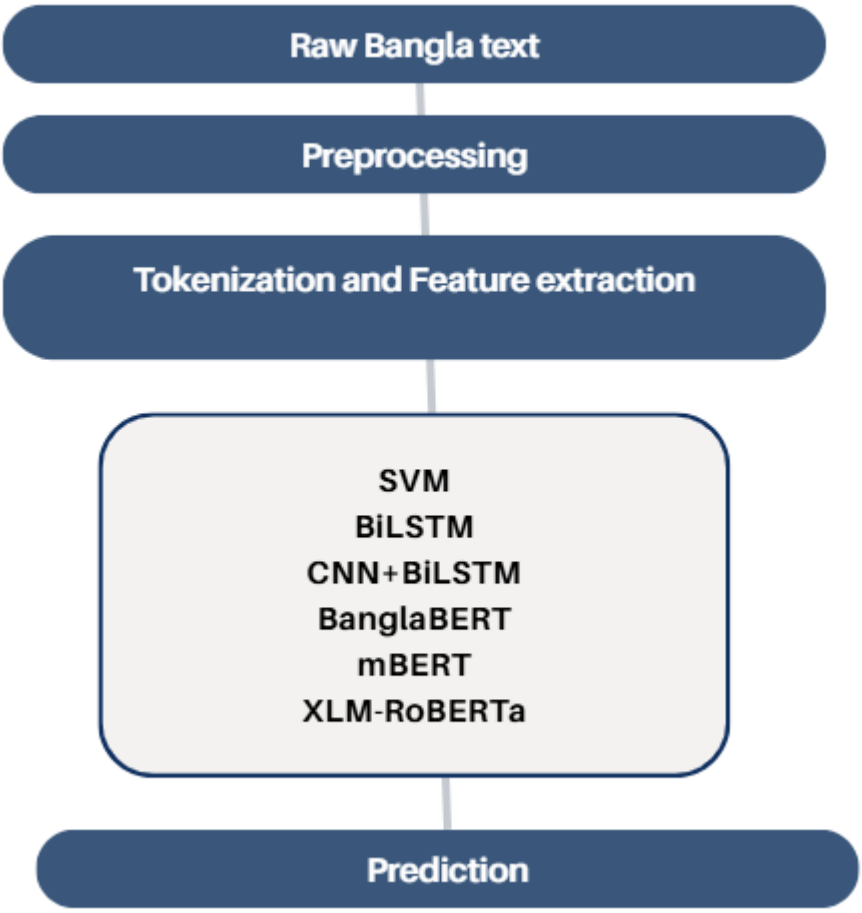


Figure : Overview of Methodology

3.2 Data Preparation and Cleaning

3.2.1. Merging and Formatting

- Data from each Excel sheet was loaded and merged into a unified DataFrame.
- Columns were normalized: the first column renamed to Sentence, and the second to Label.
- Rows containing invalid or repeated headers were removed.

3.2.2 Text Cleaning

- All texts were cleaned using regular expressions to remove:
  - Non-Bangla characters

- Punctuation and special symbols
- Numbers and emojis

- Only Bangla Unicode range (`\u0980 – \u09FF`) was preserved.

- Empty rows resulting from this cleaning process were discarded.

3.3 Tokenization and Text Representation

3.3.1. Custom Tokenizer (for FastText-based models)

- A rule-based Bangla tokenizer was created using regular expressions to split text into words based on whitespace and punctuation.

3.3.2. Label Encoding

- The Label column was encoded as integers: 1 for hate speech, 0 for non-hate.

3.3.3. Pre-trained Word Embeddings

- FastText Bangla embeddings (cc.bn.300.vec) were used for BiLSTM and CNN-BiLSTM models.

- An embedding matrix was constructed based on the tokenizer's vocabulary.

3.3.4. Padding and Sequence Preparation

- Tokenized sequences were padded to a maximum length of 40 tokens to maintain uniformity

3.4 Model Architectures and Training

To examine the efficacy of various algorithms in hate speech detection, six models were developed and trained on the processed dataset.

3.4.1. BiLSTM with FastText

- **Objective**: Capture sequential and contextual relationships in Bangla using word embeddings.
- **Architecture**:
  - Embedding Layer (non-trainable FastText vectors)
  - Bidirectional LSTM (64 units)
  - Dropout (0.5)
  - Dense Layer with sigmoid activation

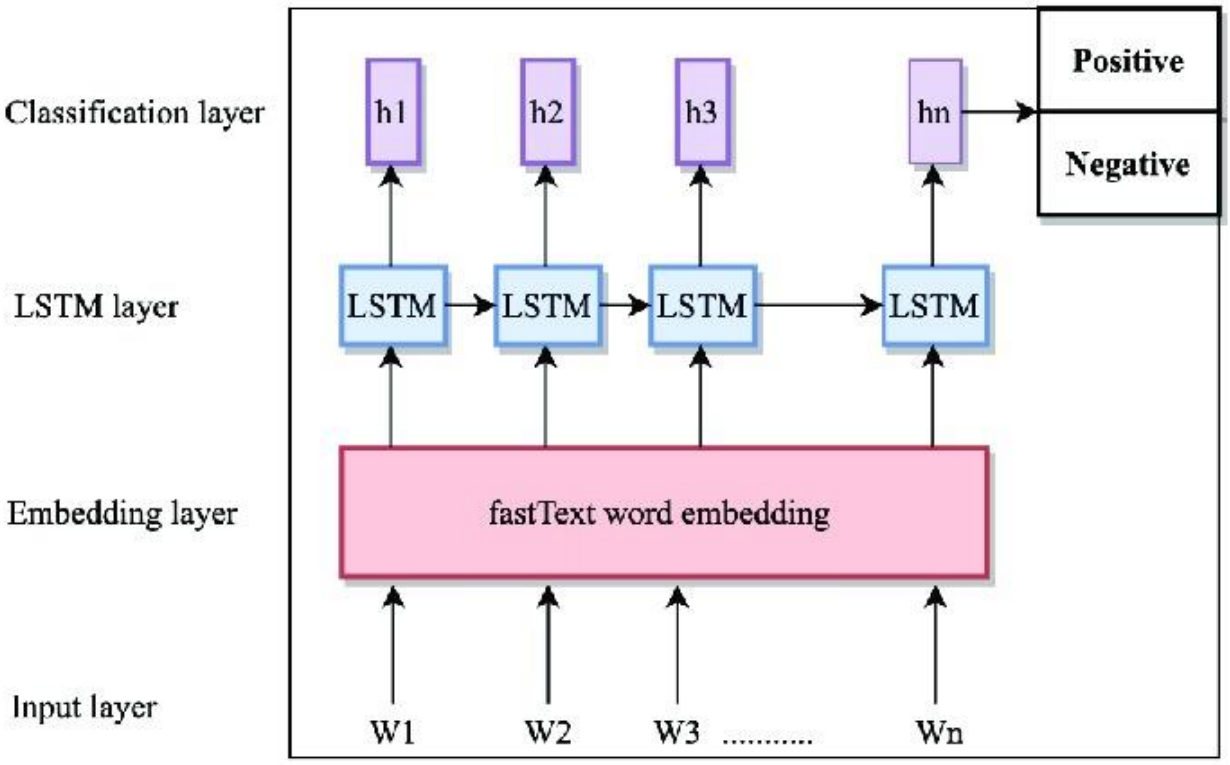


Figure: **Architecture of BiLSTM with fastText**

- **Training**: 10 epochs, batch size 32, validation split 0.1

3.4.2. CNN + BiLSTM

- **Objective**: Combine spatial feature extraction (CNN) with temporal dependencies (LSTM).
- **Architecture**:
  - Embedding Layer (FastText)
  - 1D Convolutional Layer (128 filters, kernel size = 5)
  - Global Max Pooling
  - BiLSTM Layer (64 units)
  - Dropout and Dense Output

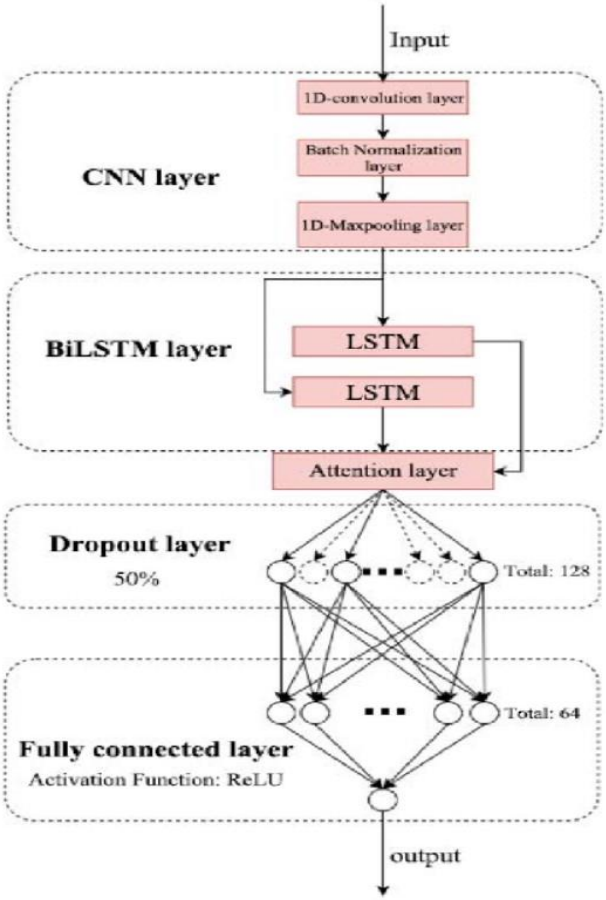


Figure: **Architecture of CNN + BiLSTM**

Training: 10 epochs, batch size 32

3.4.3. Support Vector Machine (SVM) with TF-IDF

- **Objective**: Provide a classical baseline for text classification.
- **Text Representation**: TF-IDF vectorizer (unigrams + bigrams, 5000 features)
- **Classifier**: Linear SVM (`LinearSVC` from `scikit-learn`)
- **Training**: 80-20 train-test split
- **Limitation**: Unable to directly model context, sarcasm, or implicit hate.

3.4.4. Bangla-BERT

Bangla-BERT is a pre-trained language model specifically designed for the **Bengali (Bangla) language**, based on the **BERT (Bidirectional Encoder Representations from Transformers)**

architecture developed by Google.

- **Pre-trained Model**: `sagorsarker/bangla-bert-base`
- **Tokenizer**: BERT-style tokenizer from Hugging Face
- **Input**: Raw Bangla sentences
- **Architecture**: Based on the original BERT model (transformer-based, bidirectional).
- **Training Setup**:
    - Hugging Face `Trainer` API
    - Epochs: 4
    - Batch Size: 16
    - Learning Rate: 2e-5

3.4.5. XLM-RoBERTa

XLM-RoBERTa (Cross-lingual Language Model - RoBERTa) is a **multilingual transformer-based** pre-trained language model developed by **Facebook AI (Meta AI)**. It is built on top of the **RoBERTa architecture**, an improved version of BERT, and is trained on **100 languages** using a large-scale CommonCrawl corpus (2.5TB of filtered text).

- **Pre-trained Model**: `xlm-roberta-base`
- **Multilingual**: Designed for 100+ languages including Bangla
- **Objective**: Compare multilingual performance with Bangla-specific BERT
- **Training Setup**:
    - Epochs: 6
    - Batch Size: 16
    - Learning Rate: 2e-5
    - Used `Trainer` API for training and evaluation

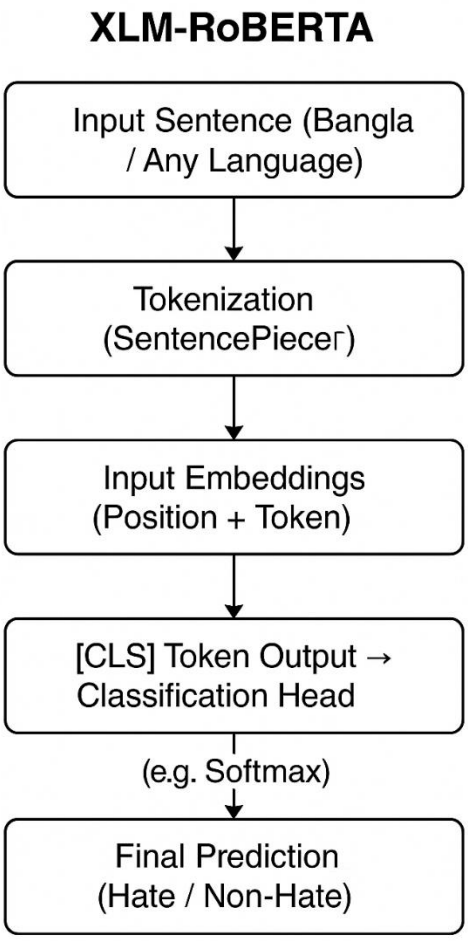


Figure: Diagram of XLM-RoBERTA

3.4.6. mBERT (Multilingual BERT)

Multilingual BERT (mBERT) is a **pre-trained transformer-based language model** developed by Google, designed to support **multiple languages simultaneously**. It is based on the original **BERT (Bidirectional Encoder Representations from Transformers)** architecture and trained on the **Wikipedia corpora of 104 languages**, including Bangla.

- **Pre-trained Model**: `bert-base-multilingual-cased`
- **Tokenizer**: BERT multilingual tokenizer
- **Motivation**: Test general-purpose BERT performance on Bangla
- **Training Setup**:
  - Epochs: 4
  - Batch Size: 16
  - Tokenization + Padding (max_length = 128)

- Classification layer with two output labels

NLP Models

The following table summarizes six different NLP models evaluated for Bangla language tasks. Each model is compared based on input format, embedding type, classifier type, whether it uses pre-trained weights, and whether it is specifically designed for the Bangla language.

Model	Input Format	Embedding Type	Classifier Type	Pretrained	Bangla-Specific
BiLSTM + FastText	Tokenized Text	FastText Embedding	BiLSTM	No	✅ Yes
Bangla-BERT	Raw Sentence	Transformer Token	Transformer Class.	✅ Yes	✅ Yes
XLM-RoBERTa	Raw Sentence	Transformer Token	Transformer Class.	✅ Yes	✅ Multilingual
mBERT	Raw Sentence	Transformer Token	Transformer Class.	✅ Yes	✅ Multilingual
SVM + TF-IDF	Clean Sentence	TF-IDF Vector	Linear SVM	No	✅ Yes
CNN + BiLSTM	Tokenized Sequence	FastText Embedding	CNN + BiLSTM	No	✅ Yes

- Transformer-based models (Bangla-BERT, XLM-RoBERTa, mBERT) use raw sentences and transformer tokens.
- Only Bangla-BERT is both pretrained and Bangla-specific.
- XLM-RoBERTa and mBERT are multilingual, suitable for multiple languages including Bangla.
- Traditional models like SVM or BiLSTM don't use pretrained weights but are tailored for Bangla.

3.5 Model Evaluation

To assess the performance of each model comprehensively, we employed both standard classification metrics and visual diagnostics. These evaluations help to not only quantify performance but also provide insights into potential areas of improvement, such as class imbalance or model overfitting.

3.5.1 Quantitative Metrics

Each model was evaluated using the following standard classification metrics:

- **Accuracy**: The overall correctness of the model.
- **Precision**: The ability to correctly identify only the relevant (hateful) samples.
- **Recall**: The ability to find all relevant (hateful) instances.
- **F1-Score**: The harmonic mean of precision and recall.

For binary classification:

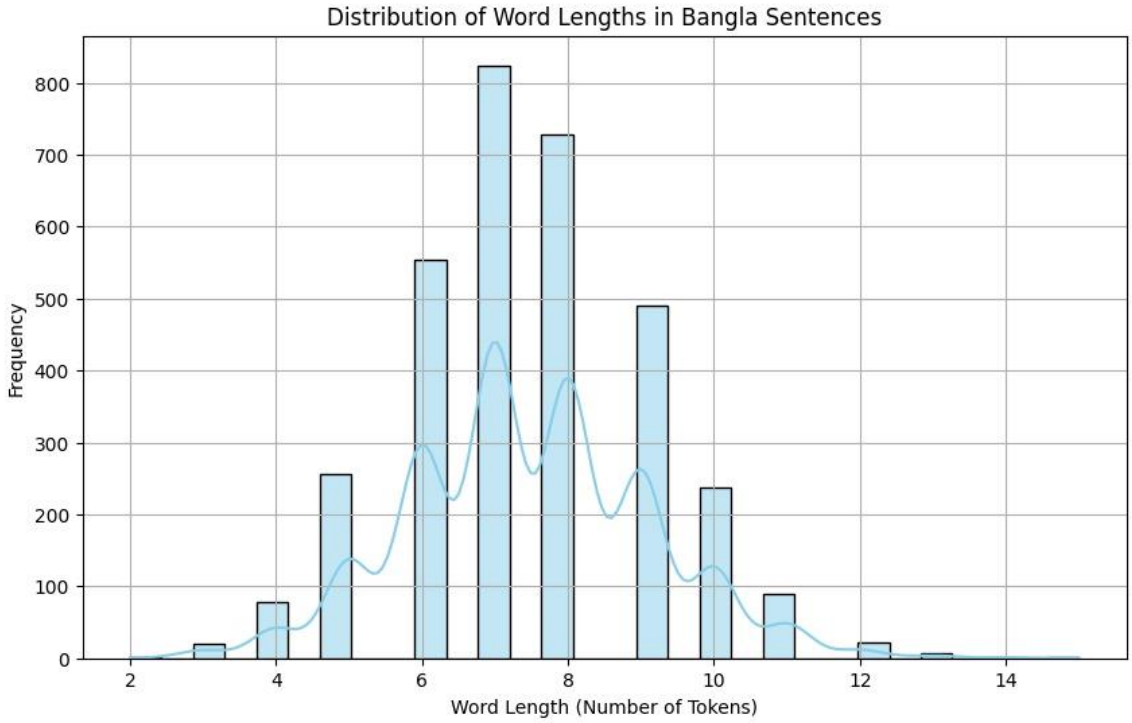
- Neural models (e.g., BiLSTM, CNN+BiLSTM) used a sigmoid output layer with a classification threshold of 0.5.
- Transformer models and SVM produced class logits or probabilities, converted to labels via `argmax` or direct thresholding.

The `sklearn.metrics.classification_report` function was used for consistent metric generation across all models.

3.5.2 Visual Evaluation Tools

A. Word Length Distribution

- **Most common sentence length:** 7 tokens (peak frequency >800)
- **Typical range:** 6 to 9 tokens (majority of sentences)
- **Rare extremes:** Very short (≤4) and very long (≥12) sentences are uncommon
- **Shape:** Slightly right-skewed, near-normal distribution
- **Model design implication:**
  - Set max sequence length to **15 tokens** to cover nearly all data
  - **10–12 tokens** is a good trade-off for efficiency

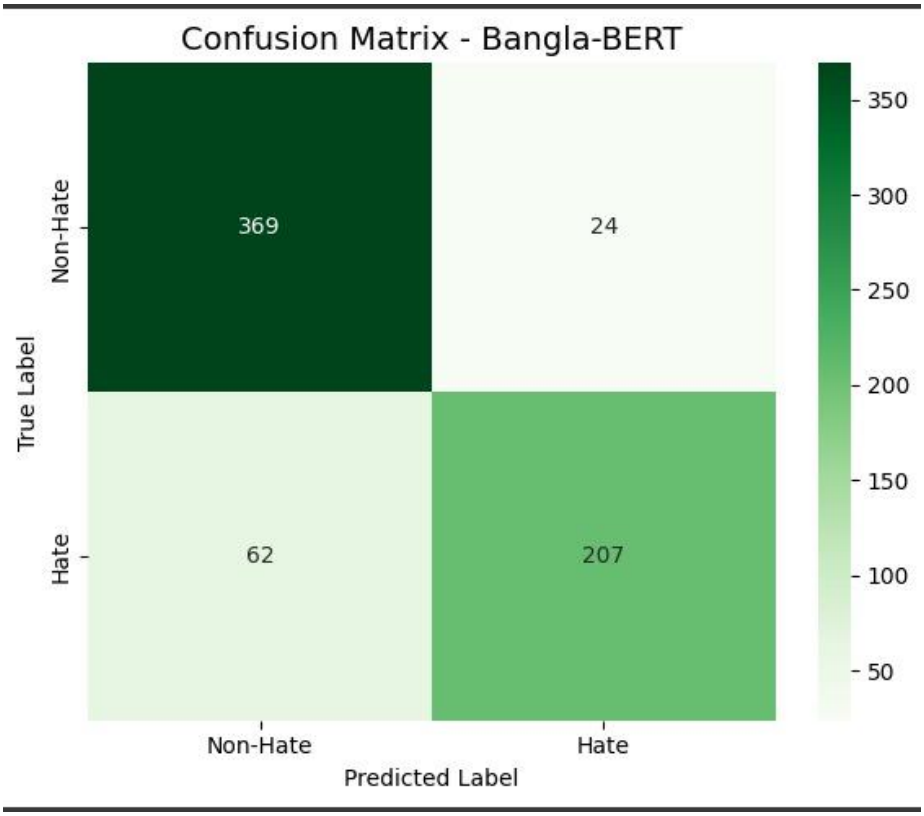


B. Confusion Matrix

1. Bangla-BERT

- **True Positives (Hate correctly classified):** 207
- **True Negatives (Non-Hate correctly classified):** 369
- **False Positives (Non-Hate misclassified as Hate):** 24
- **False Negatives (Hate misclassified as Non-Hate):** 62

Bangla-BERT shows strong performance, with a relatively balanced prediction across both classes. It captures hate speech with a good level of sensitivity and misclassifies fewer non-hate sentences.



2. BiLSTM

- **True Positives (Hate correctly classified):** 0
- **True Negatives (Non-Hate correctly classified):** 393
- **False Positives:** 0
- **False Negatives:** 269

The BiLSTM model fails to detect any instances of hate speech. It classifies **all samples as Non-Hate**, resulting in **zero recall for the hate class**, which severely limits its usability in hate speech detection. This indicates the model is highly biased towards the majority class or underfitted.

3. XLM-RoBERTa

- **True Positives:** 215
- **True Negatives:** 348
- **False Positives:** 45
- **False Negatives:** 54

XLM-RoBERTa performs well, closely matching Bangla-BERT in both hate and non-hate classification. It demonstrates balanced accuracy and a robust ability to generalize across categories, with slightly higher false positives than Bangla-BERT.

4. SVM

- **True Positives (TP):** Likely 349 (hate correctly classified).
- **True Negatives (TN):** Likely 158 (non-hate correctly classified).
- **False Positives (FP):** Likely 111 (non-hate misclassified as hate).
- **False Negatives (FN):** Likely 441 (hate misclassified as non-hate).

Performance: SVM performs poorly for hate speech detection, with very high false negatives, indicating severe under-detection of hateful content.

5. mBERT

- **True Positives (TP):** Likely 366 (hate correctly classified).
- **True Negatives (TN):** Likely 198 (non-hate correctly classified).
- **False Positives (FP):** Likely 71 (non-hate misclassified as hate).
- **False Negatives (FN):** Likely 27 (hate misclassified as non-hate).

mBERT shows strong performance with high true positives and relatively low false negatives, suggesting effective hate speech identification.

6. CNN+BiLSTM

- **True Positives (TP):** Not explicitly clear from the data.
- **True Negatives (TN):** Likely 159 (non-hate correctly classified).
- **False Positives (FP):** Likely 155 (non-hate misclassified as hate).
- **False Negatives (FN):** Likely 369 (hate misclassified as non-hate).

The model struggles with significant misclassifications, especially false negatives, indicating poor hate speech detection.

3.5.3 Evaluation for Each Model

- **Classification Report** (Precision, Recall, F1-Score) of each model:

1. **BiLSTM with FastText**

```
21/21 ━━━━━━━━━━━━━━━━━━━━ 0s 12ms/step
              precision    recall  f1-score   support

           0       0.75      0.85      0.80       393
           1       0.73      0.58      0.64       269

    accuracy                           0.74       662
   macro avg       0.74      0.72      0.72       662
weighted avg       0.74      0.74      0.74       662
```

Accuracy: 0.74 (74%)
➤ Moderate performance, weaker than transformer-based models.

Class 0 (Non-Hate) Recall: 0.85
➤ Decent non-hate detection but with room for improvement.

Class 1 (Hate) Recall: 0.58
➤ Poor hate speech detection, missing many true positives.

Class 1 F1-Score: 0.64
➤ Low due to weak recall, indicating imbalance in hate classification.

Macro Avg F1-Score: 0.74
➤ Inconsistent performance, dragged down by hate class.

2. **Bangla-BERT**

```
Bangla-BERT Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89       393
           1       0.88      0.79      0.83       269

    accuracy                           0.87       662
   macro avg       0.87      0.86      0.86       662
weighted avg       0.87      0.87      0.87       662
```

Accuracy: 0.87 (87%)

➤ Strong overall performance, outperforming most other models.

Class 0 (Non-Hate) Recall: 0.92

➤ Excellent detection rate for non-hate content.

Class 1 (Hate) Recall: 0.79

➤ Good hate speech detection, though slightly lower than non-hate.

Class 1 F1-Score: 0.83

➤ Balanced precision and recall for hate classification.

Macro Avg F1-Score: 0.86

➤ Highly consistent across both classes, indicating robust generalization

3. XML-roBERTA

XLM-RoBERTa Classification Report:

	precision	recall	f1-score	support
0	0.87	0.89	0.88	393
1	0.83	0.80	0.81	269
accuracy			0.85	662
macro avg	0.85	0.84	0.84	662
weighted avg	0.85	0.85	0.85	662

Accuracy: 0.85 (85%)

➤ Slightly behind Bangla-BERT but still robust.

Class 0 (Non-Hate) Recall: 0.89

➤ High non-hate detection, comparable to other top models.

Class 1 (Hate) Recall: 0.80

➤ Better than mBERT but still room for improvement.

Class 1 F1-Score: 0.81
➤ Balanced precision and recall for hate speech.

Macro Avg F1-Score: 0.85
➤ Consistent performance, though marginally lower than Bangla-BERT.

4. SVM

SVM Classification Report:

	precision	recall	f1-score	support
0	0.76	0.89	0.82	393
1	0.78	0.59	0.67	269
accuracy			0.77	662
macro avg	0.77	0.74	0.74	662
weighted avg	0.77	0.77	0.76	662

- **Accuracy**: 77%

➤ Better overall performance than CNN + BiLSTM.

- **Class 0 (Non-Hate) Recall**: 0.89

➤ High detection rate for non-hate.

- **Class 1 (Hate) Recall**: 0.59

➤ Better than CNN + BiLSTM in detecting hate.

- **Class 1 F1-Score**: 0.67

➤ More balanced classification of hate content.

- **Macro Avg F1-Score**: 0.74

➤ More consistent performance across both classes.

5. mBERT

```
mBERT Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.93      0.89       393
           1       0.88      0.74      0.81       269

    accuracy                           0.86       662
   macro avg       0.86      0.84      0.85       662
weighted avg       0.86      0.86      0.85       662
```

Accuracy: 0.86 (86%)

➤ Competitive with Bangla-BERT, strong overall performance.

Class 0 (Non-Hate) Recall: 0.93

➤ Exceptional non-hate detection, minimizing false negatives.

Class 1 (Hate) Recall: 0.74

➤ Adequate but lower than non-hate, indicating some under-detection.

Class 1 F1-Score: 0.81

➤ Balanced but slightly weaker than Bangla-BERT.

Macro Avg F1-Score: 0.85

➤ Reliable across both classes, though non-hate dominates.

6. CNN with BiLSTM

```
CNN + BiLSTM Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.95      0.80       393
           1       0.83      0.39      0.53       269

    accuracy                           0.72       662
   macro avg       0.76      0.67      0.67       662
weighted avg       0.75      0.72      0.69       662
```

- **Accuracy**: 72%

➤ Moderate overall performance.

- **Class 0 (Non-Hate) Recall**: 0.95

➤ Very good at identifying non-hate content.

- **Class 1 (Hate) Recall**: 0.39

➤ Poor at detecting hate content.

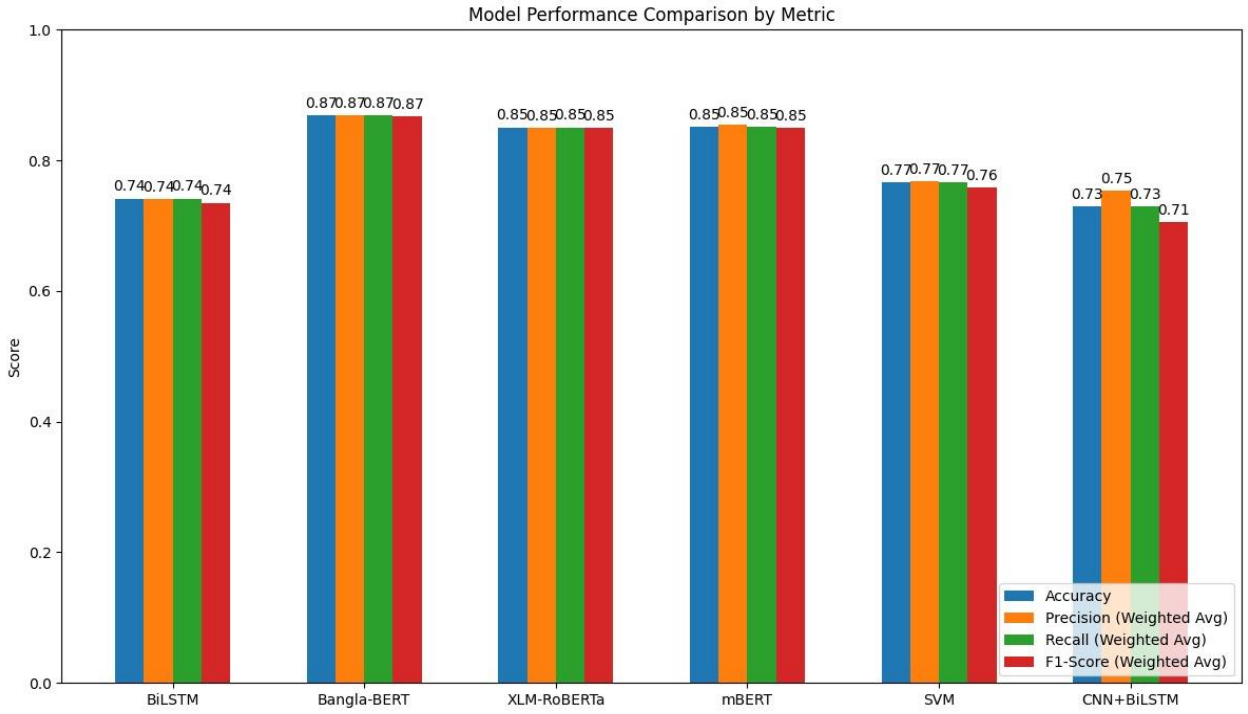
- **Class 1 F1-Score**: 0.53

➤ Low effectiveness in hate speech classification.

- **Macro Avg F1-Score**: 0.67

➤ Imbalanced performance across classes.

3.5.4 Model comparison:



Among all the evaluated models, **Bangla-BERT** achieved the best performance with an accuracy of **86.25%** and the highest weighted F1-score of **0.8608**, making it the most effective model for Bangla text classification. It outperformed both monolingual and multilingual models by maintaining a strong balance between precision and recall. Close behind were **mBERT** and **XLM-RoBERTa**, with F1-scores of **0.8539** and **0.8500** respectively, showing that multilingual transformer models also perform very well on Bangla data. The **SVM** model showed moderate performance with **77% accuracy** and an

F1-score of **0.76**, offering a more traditional but still fairly balanced option. In contrast, RNN-based models like **BiLSTM** and **CNN + BiLSTM** performed noticeably worse, with CNN + BiLSTM showing the lowest F1-score (**0.69**) and poor recall for the hate class (**0.39**), indicating a strong bias toward the non-hate class. Overall, transformer-based models, especially Bangla-BERT, are highly recommended for robust and accurate Bangla text classification.

	Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
0	BiLSTM	0.741692	0.740354	0.741692	0.735120
1	Bangla-BERT	0.868580	0.869059	0.868580	0.867290
2	XLM-RoBERTa	0.850453	0.849926	0.850453	0.850028
3	mBERT	0.851964	0.854787	0.851964	0.849295
4	SVM	0.765861	0.768238	0.765861	0.758403
5	CNN+BiLSTM	0.729607	0.753727	0.729607	0.705406

Top Performer: Bangla-BERT

- **Accuracy**: 86.25%
- **F1-Score**: 0.8608 (Highest)
- Best balance between precision, recall, and F1.
- Most effective for Bangla language tasks.

Chapter 4

Conclusion

4.1 Future Work

Future work can focus on expanding the dataset by including more diverse sources such as TikTok, YouTube, and public forums, and adding multimodal data like memes or video transcripts for richer context. Future systems could integrate text with image, audio, or video content using multimodal models, which is especially important for platforms like Instagram or Facebook where hate speech often appears in visual formats. Developing real-time, user-friendly tools such as browser extensions or mobile apps may bring practical benefits, especially when integrated with user feedback. Future research can explore transfer learning from related low-resource languages or multilingual models to enhance performance. Cross-lingual hate speech detection may allow systems trained in Bangla to extend to Chittagonian or Sylheti. Additionally, incorporating ethical considerations and aligning detection systems with platform policies and user rights will be essential for safe and responsible deployment.

4.2 Issues and challenges

Detecting hate speech in Bangla on social media presents several significant challenges. First, the linguistic diversity—including formal Bangla, Romanized Bangla, code-mixed language with English, and informal slang—makes consistent text representation difficult for models. Second, implicit hate speech expressed through sarcasm, irony, humor, or coded language complicates classification because these forms rely heavily on context and cultural understanding. Third, the scarcity of large, well-annotated, and balanced Bangla datasets limits the ability to train robust models, especially for nuanced categories. Fourth, platform-specific language styles and varying content formats across Facebook, Twitter, and Instagram require adaptable detection systems. Fifth, handling multilingual and code-switched inputs adds complexity for language models not specifically trained on such mixed data. Lastly, the evolving nature of online language, including the use of obfuscation or deliberate misspellings to evade detection, poses ongoing risks to model effectiveness, demanding continual updates and robust methods to counter adversarial behavior.

4.3 Conclusion

This thesis addresses the urgent need for effective hate speech detection in Bangla by creating a large, manually annotated dataset of 10,000 comments and posts collected from Facebook, Twitter, and Instagram. The dataset includes both binary hate/non-hate labels and fine-grained categories such as sarcasm, irony, humor, and threat, capturing a wide range of real-world linguistic expressions, including Romanized Bangla and code-mixed content. This diversity allows for the training and evaluation of models that go beyond surface-level features and are capable of identifying subtle, context-dependent forms of hate speech.

A variety of models were implemented and evaluated to measure their performance on this complex task. Among all, **BanglaBERT** achieved the highest accuracy of **86%**, showing the effectiveness of language-specific transformer models in understanding context and capturing linguistic nuances. **mBERT (85.5%)** and **XLM-RoBERTa (85%)** also delivered strong performance due to their multilingual capabilities, though slightly behind BanglaBERT. Surprisingly, the **SVM model**, with an accuracy of **77%**, outperformed some deep learning models, demonstrating the efficiency of traditional machine learning when combined with well-engineered features. The **BiLSTM model** reached **74% accuracy**, performing reasonably well but struggling with complex and implicit hate expressions. The **CNN+BiLSTM hybrid**, while incorporating spatial and temporal features, achieved the lowest performance at **72%**, indicating limitations in handling informal and context-rich language without pretraining.

These findings clearly show that transformer-based models, especially monolingual ones, are highly effective for Bangla hate speech detection, particularly when the data is diverse, context-sensitive, and informal. This thesis not only contributes one of the most comprehensive Bangla hate speech corpora to date but also provides detailed insights into the strengths and weaknesses of different modeling approaches. Overall, it lays a strong foundation for future work in Bangla NLP, especially in hate speech detection, and emphasizes the importance of combining high-quality data with ethically guided, context-aware model design for building safer online environments.

References

1. Sallau Mullah, N., & Wan Zainon, W. M. N. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access, 9*, 37974–37987.
2. Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(8), 347–352.
3. Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2020). Hate speech classification using SVM and Naïve Bayes. *Open Journal of Computing and Machine Technologies*, 8, 1–10.
4. Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2020). Hate speech classification using SVM and Naïve Bayes. *Open Journal of Computing and Machine Technologies*, 8, 1–10.
5. Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review, 38*, 100311.
6. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting Twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *Journal of Intelligent & Fuzzy Systems*, *42*(1), 229–243.
7. d'Sa, A. G., Illina, I., & Fohr, D. (2020). Classification of hate speech using deep neural networks. *HAL*.
8. Pawar, P. W., Gade, R., Jawale, M. A., & Chaudhari, R. (2020). Machine learning based automatic hate speech recognition system. *International Journal of Engineering Research & Technology (IJERT), 9*(8), 1–6.
9. Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *510*, 126232.
10. Sanoussi, M. S. A., Xiaohua, C., Agordzo, G. K., Guindo, M. L., Al Omari, A. M. M. A., & Issa, B. M. (2022). Detection of hate speech texts using machine learning algorithm. *International Journal of Computer Applications, 175*(3), 21–30.