Port City International University
Excellence in Higher Education
UGC & Govt. Approved University at Khulshi in Chittagong

Developing a Dataset for Toxicity Detection in Bangla-English Code-Mixed in Social Media

---

Supervised By :

Tanvinur Rahman Siam

Lecturer
Department of CSE
Port City International University.

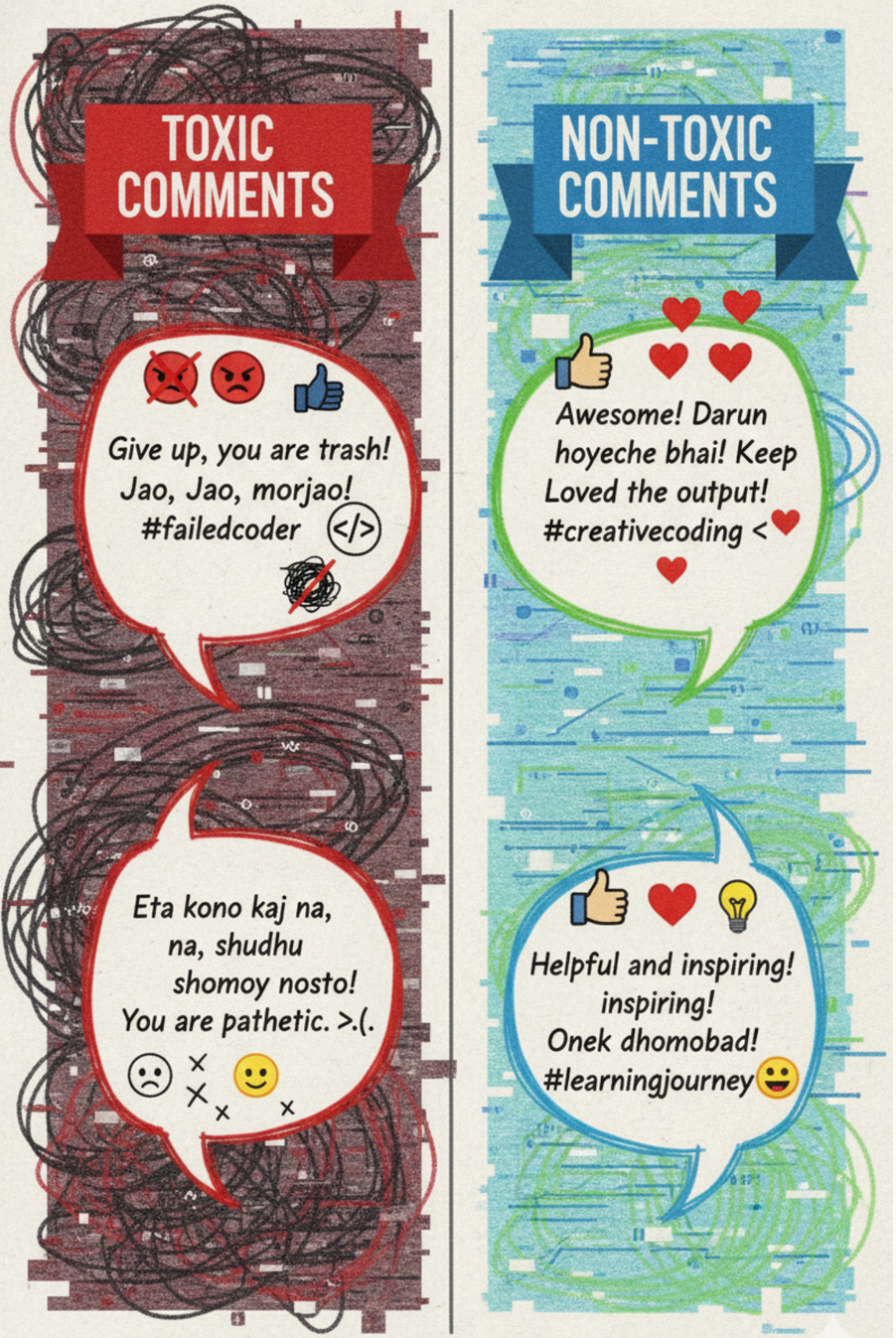
Proposed By :

Shaik Abdul Ahad

ID: CSE 029 07666
Batch: CSE 29-E-A
Port City International University

Department Of Computer Science and Engineering
Port City International University
Proposal Fall 25

List of Contents



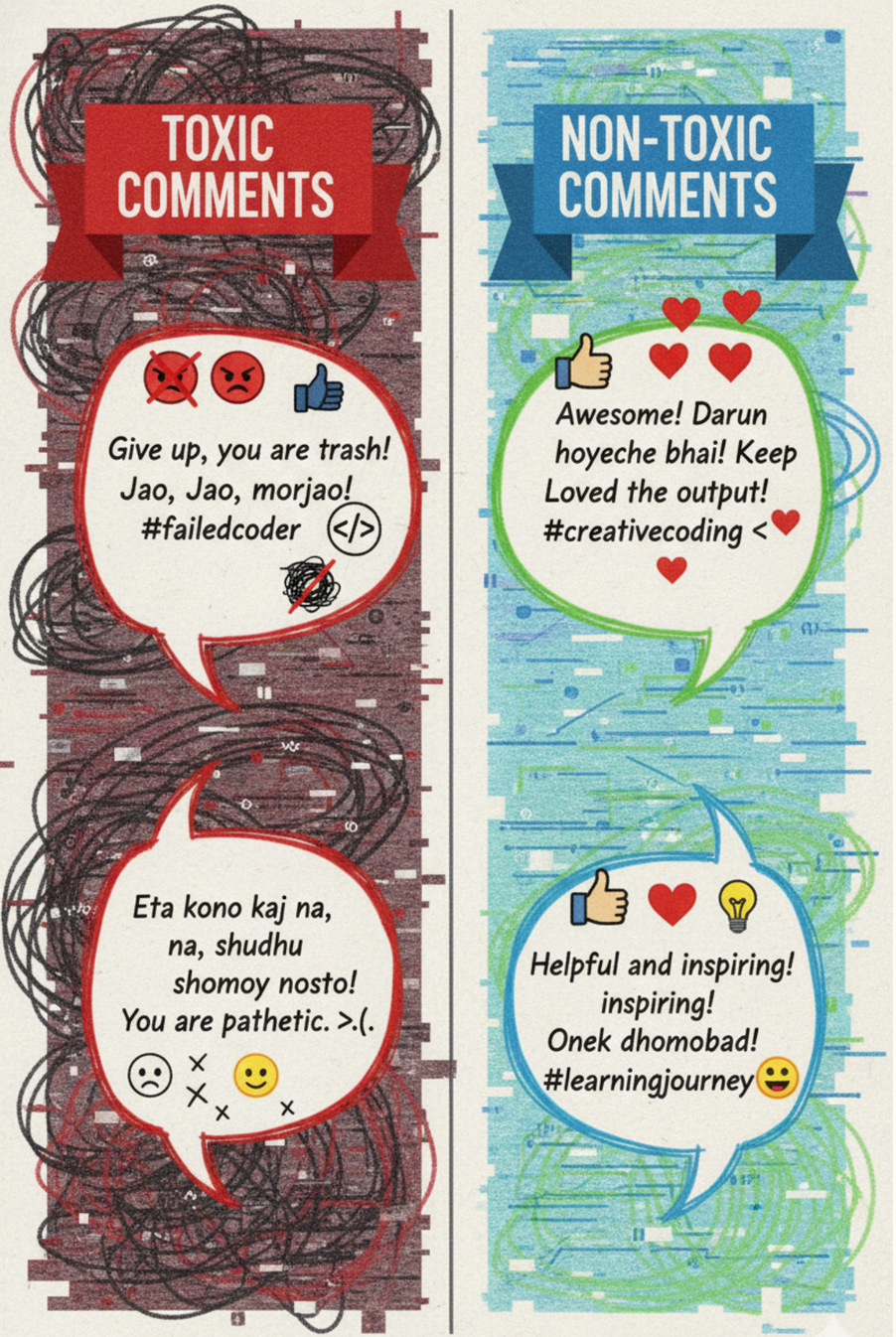
Source: AI-Generated Image

Introduction

The Rise of Code-Mixed Communication

- Bangladesh has over 52.90 million active social media users (DataReportal, 2024).
- Bangla-English code-mixing is now a defining feature of social media discourse.
  - Example: “Tumi kemon acho? I’m fine.”
- Romanized Bangla adds another layer of complexity due to inconsistent spelling.

Source: DataReportal, 2024
Accessed Date: 07.10.2025

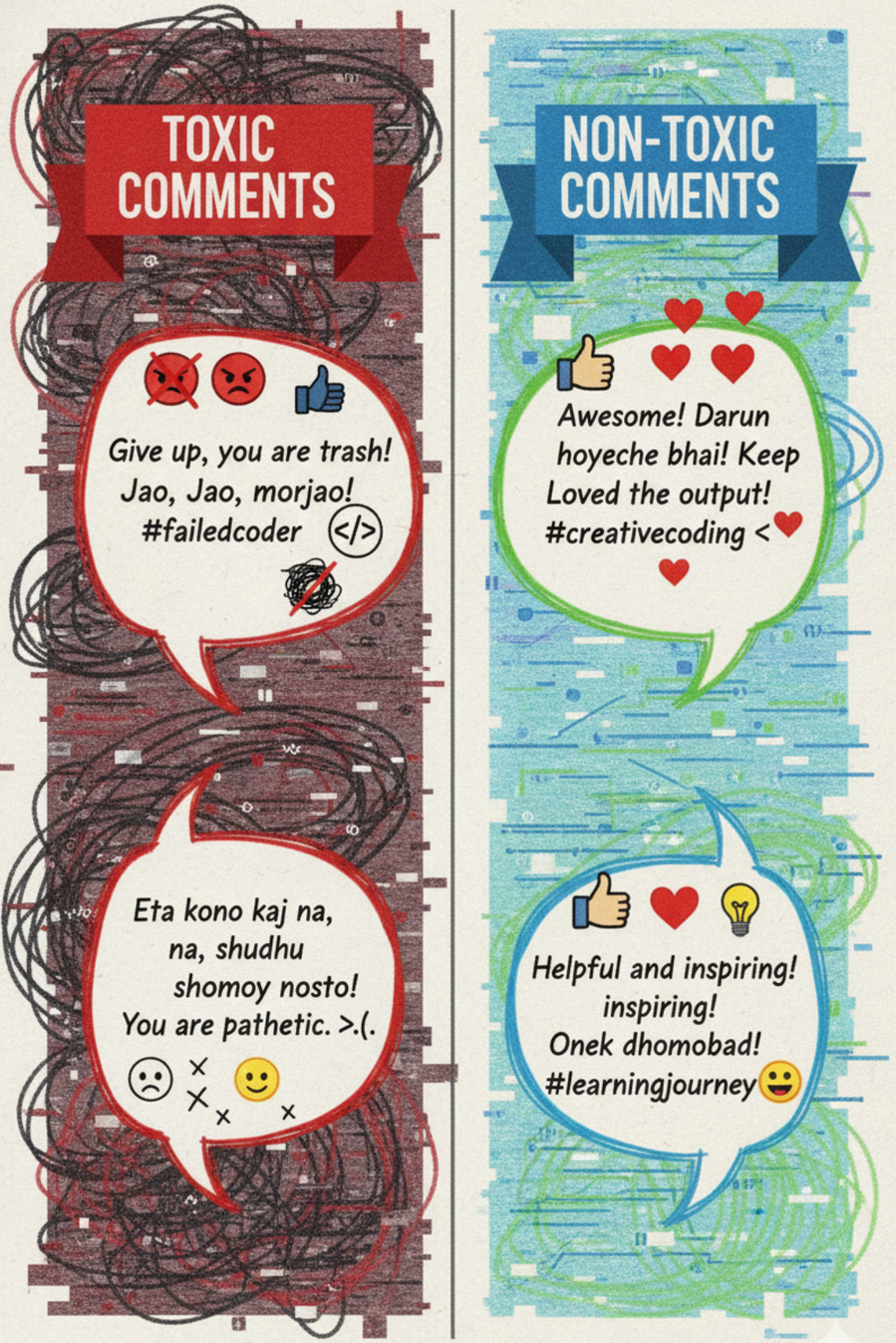


Source: AI-Generated Image

Introduction(Cont.)

The Problem

- Toxic language spreads rapidly online, harming users and communities.
- Existing detection systems fail to handle Bangla-English code-mixed text.
- Reasons:
  - Models trained mostly on monolingual data (English or Bangla).
  - Cultural nuances and script variations remain underrepresented.



Source: AI-Generated Image

Introduction(Cont.)

What is Toxic Content?

- Toxic content includes insults, threats, hate speech, and harassment based on religion, race, gender, or ethnicity.
- It differs from simple disagreement or criticism.

What is Code-Mixing?

- Code-mixing involves combining two or more languages in a single utterance.
- Examples in Bangla-English:
  - "Tumi kemon acho? I'm fine."
  - "aj er match ta darun chilo bro"
  - "ei post ta completely faltu"

Goal

Overarching Aim

- To create the first publicly available benchmark dataset for Bangla-English code-mixed toxicity detection, and establish baseline NLP models for this task.

Specific Objectives

- Collect, clean, and annotate Bangla-English social media comments.
- Implement and evaluate traditional ML and transformer-based models.
- Publish dataset, guidelines, and source code for public research use.

Motivation and Application

MOTIVATION

- Code-mixed toxicity detection is a neglected research area.
- Lack of standard datasets and annotation frameworks for Bangla-English text.
- Growing need for culturally aware NLP systems in South Asia.

APPLICATIONS

- Automatic filtering of toxic comments on Facebook, YouTube, and X (Twitter).
- Cyberbullying prevention and safer online communities.
- Market analytics and sentiment monitoring in Bangladeshi contexts.
- Benchmark for future low-resource and code-mixed NLP research.

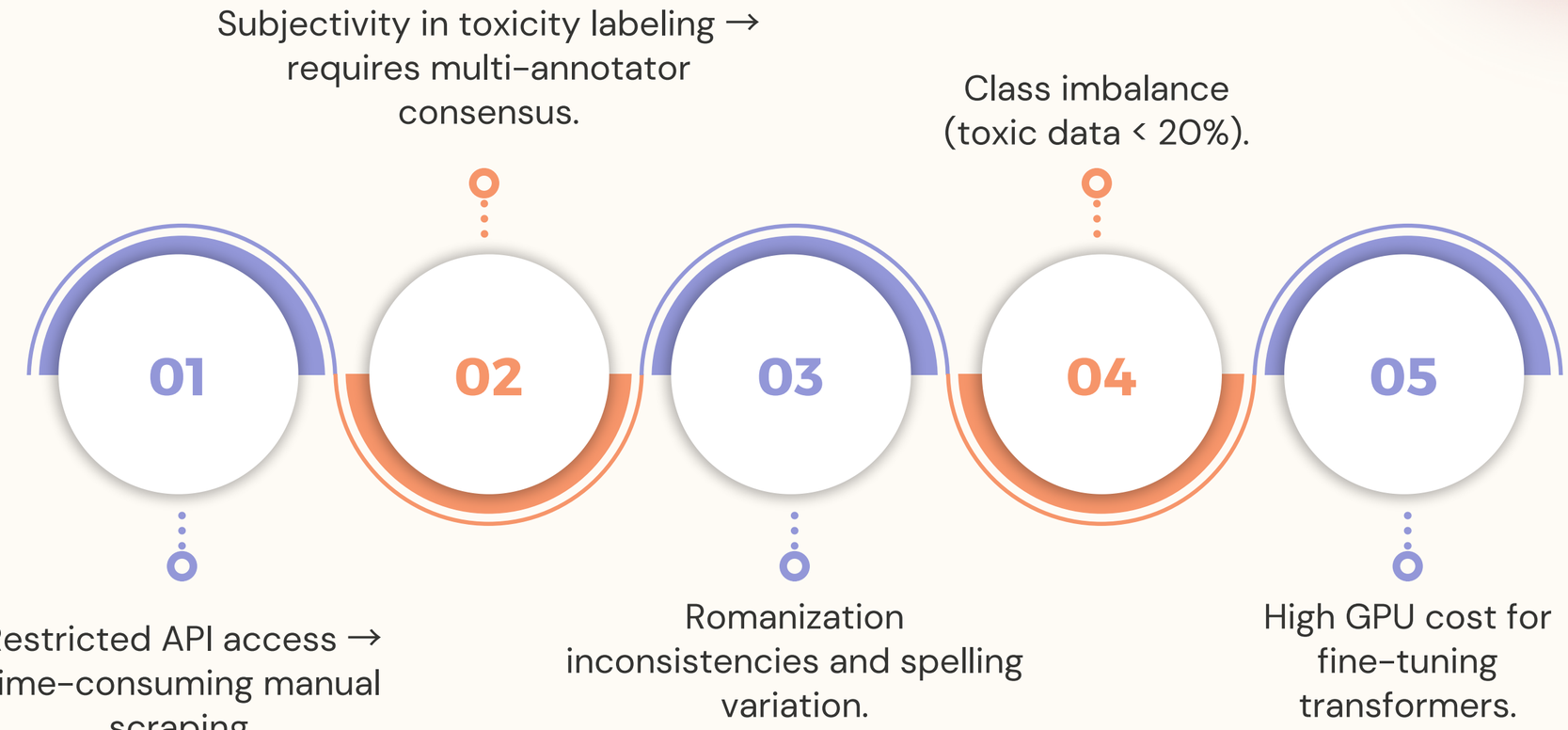
Related Work

SL	Author & Year	Publisher	Dataset	Model	Result	Observations
1	Raihan et al. (2023)	ACL (BanglaNLP)	TB-OLID (5k Bangla-English)	fBERT, HateBERT, mBERT, XLM-R, etc.	Best models (fBERT, HateBERT)	English pre-trained transformers perform best on transliterated + code-mixed Bangla
2	Goswami et al. (2023)	ACL (SocialNLP)	OffMix-3L (Bangla-English-Hindi)	BanglishBERT, HateBERT, mBERT, XLM-R	BanglishBERT = 0.68; HateBERT = 0.60; mBERT / XLM-R = 0.88 (synthetic)	BanglishBERT outperforms on real data; synthetic data yields very high performance

Related Work(Cont.)

3	Mathur et al. (2018)	ACL Workshop	Hindi-English (HEOT	Transfer CNN + LSTM	F1 = 0.714 (with transfer learning)	Transfer learning from English → Hinglish helps performance
4	Chakravarthi et al. (2020)	FIRE 2020(HASOC track)	HASOC (Multilingual: English, Hindi, German)	BERT variants, SVM, etc.	F1s around 0.51–0.53 for binary tasks; lower for fine-grained	Multilingual benchmark showing difficulty of hate / offensive classification across languages
5	Romim et al. (2021)	EACL 2021	Bengali Hate Speech (monolingual)	SVM, CNN, Deep Learning models	SVM achieved best accuracy = 87.5% on this dataset (accuracy)	Monolingual Bangla dataset; lacks code-mixed samples

CHALLENGES AND LIMITATIONS



RESEARCH GAP

- No benchmark dataset exists for Bangla-English code-mixed toxicity detection.
- Annotation guidelines for bilingual toxic content are undefined.
- Bangla-English remains understudied compared to Hindi-English.
- Multilingual transformers (XLM-R, mBERT) remain unevaluated in this context.

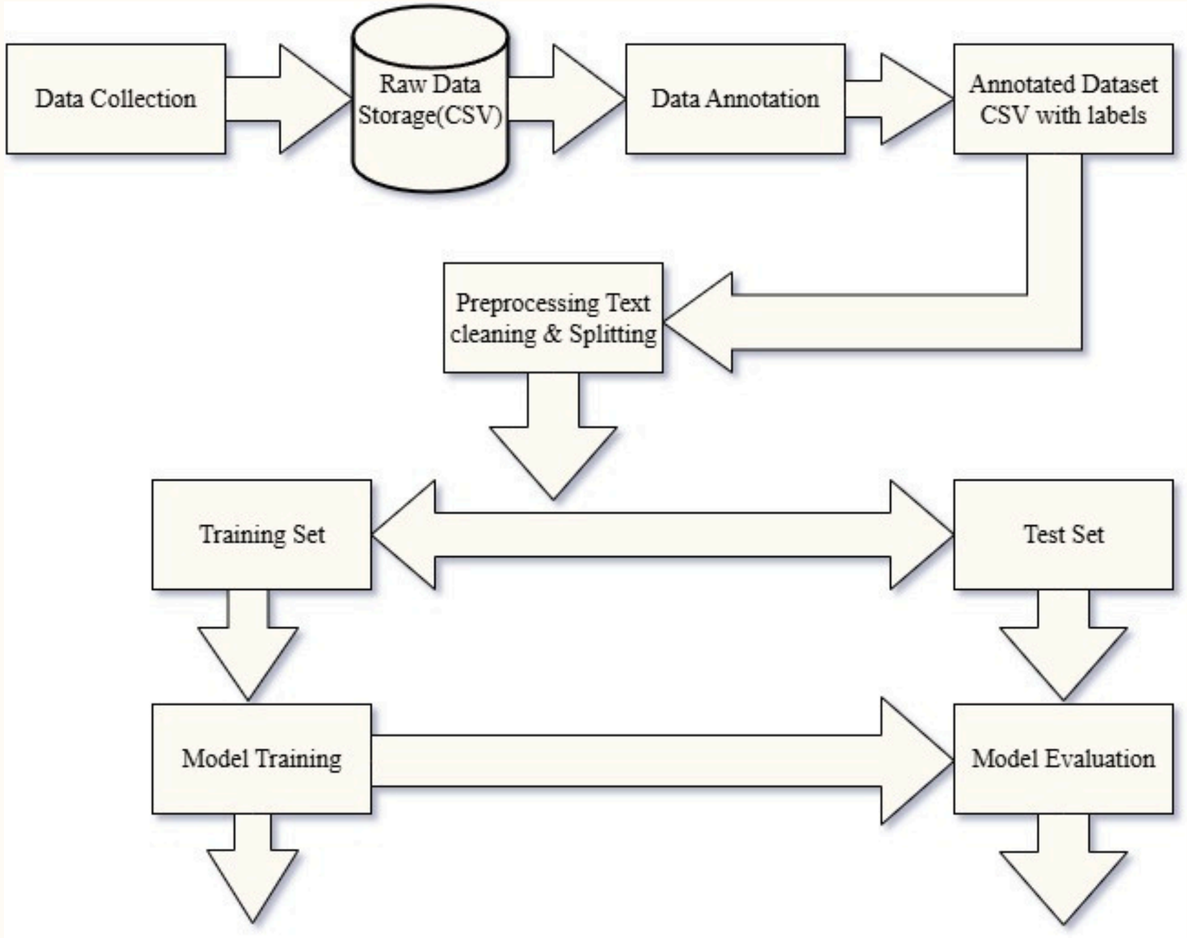
Research Questions

- Can a reliable dataset for Bangla-English code-mixed toxicity detection be built?
- How do ML models (TF-IDF + SVM) compare to transformers (XLM-R, BanglaBERT)?
- What are the failure patterns across models?

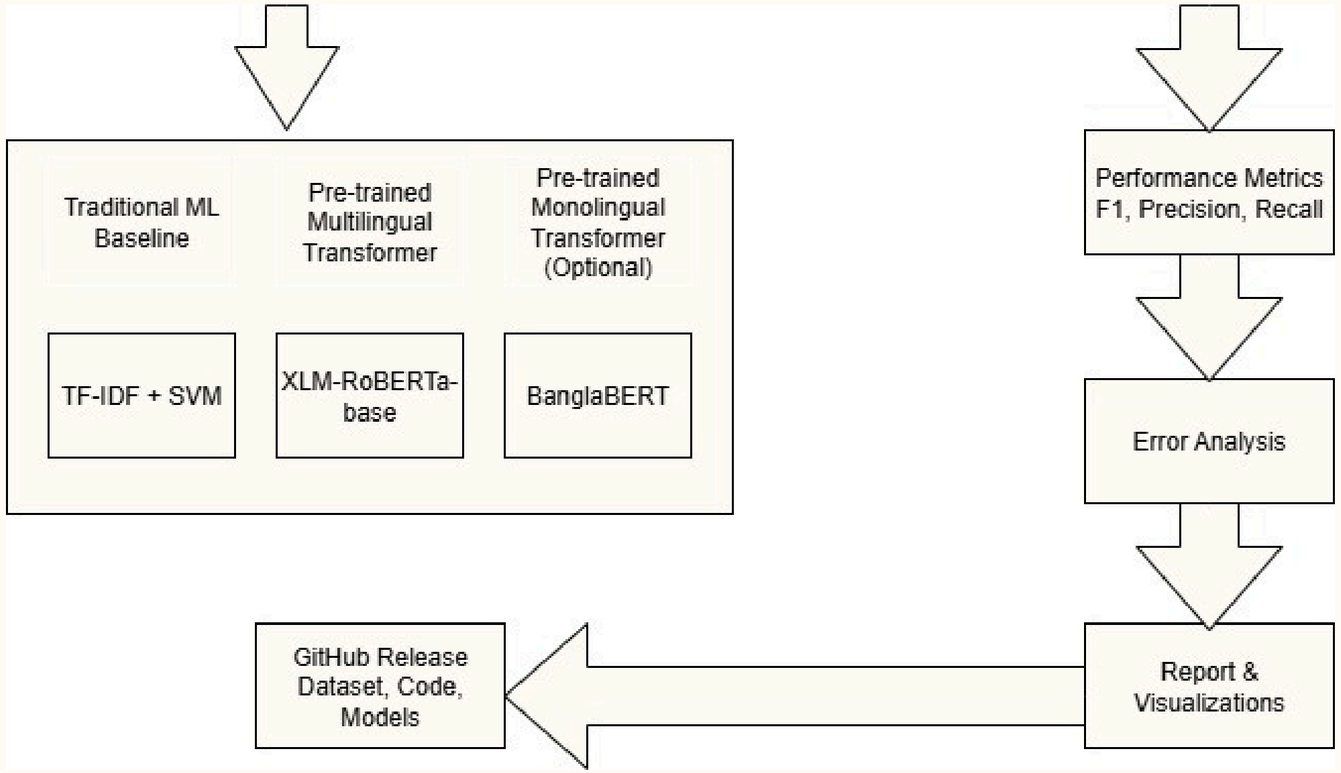
Research Objectives

- Collect 1,000–1,200 code-mixed social media comments.
- Develop clear annotation guidelines and achieve Cohen’s Kappa > 0.65.
- Preprocess text (romanization, token tagging, cleaning).
- Implement TF-IDF + SVM, XLM-RoBERTa, and BanglaBERT.
- Evaluate using F1-score, precision, recall, and confusion matrix.
- Conduct error analysis and publish dataset + code.

Proposed Methodology



Proposed Methodology (Cont.)



Requirements Analysis

Hardware

- GPU-enabled system (8–32 GB RAM, 200 GB SSD).
- Reliable internet connection.

Requirements Analysis(Cont.)

Software

- Python 3.8+
- Libraries: transformers, scikit-learn, pandas, numpy, torch
- Google Forms for annotation.
- GitHub for version control.
- Cloud platform: Google Colab

Cost Estimation

Cost Estimation Item	Description	Estimated Cost (BDT)
Data Collection	Internet, scripts, and API usage	2,000
Data Annotation	Manual labeling by 3 annotators (per 500 samples)	1,500
Data Storage & Tools	Google Drive, GitHub, cloud hosting, etc.	1,000
Model Training Resources	GPU/Colab Pro subscription	2,500

Cost Estimation(Cont.)

Software & Libraries	Python, NLP packages (open-source tools)	0
Documentation & Printing	Report printing, binding, stationery	1,000
Miscellaneous Expenses	Internet, backup, and contingency costs	1,000
Total Estimated Cost		**9,000 BDT**

Expected Impact

National Impact

- Safer online environment for Bangladeshi users.
- Supports content moderation and cyberbullying prevention.
- Strengthens local NLP research infrastructure.

International Impact

- Advances low-resource NLP and multilingual AI ethics.
- Provides a replicable framework for similar languages.

Long-Term Vision

- Foundation for future Bangla-English NLP tasks.
- Encourages open data, reproducibility, and regional innovation.

References

[1] M. N. Raihan et al., “Offensive Language Identification in Transliterated and Code-Mixed Bangla (TB-OLID),” BanglaNLP Workshop, ACL 2023, 2023. [CrossRef]

[2] D. Goswami et al., “OffMix-3L: A Code-Mixed Dataset in Bangla-English-Hindi for Offensive Language Detection,” SocialNLP Workshop, ACL 2023, 2023. [CrossRef]

[3] P. Mathur et al., “Detecting Offensive Tweets in Hindi-English Code-Switched Language,” ACL Workshop on Linguistic Code Switching, 2018. [CrossRef]

[4] B. R. Chakravarthi et al., “Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Content Identification,” FIRE Conference, 2020. [CrossRef]

[5] N. Romim et al., “Hate Speech Detection in the Bengali Language: Dataset and Baseline Results,” EACL Conference, 2021. [CrossRef]

Thank you