BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

MixSarc: A Bangla-English Code-Mixed Corpus For Implicit Meaning Identification

Tamim Ahmed
200041150
Kazi Samin Yasar Alam
200041119
Md Tanbir Chowdhury
200041114

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Tamim Ahmed**, **Kazi Samin Yasar Alam**, and **Md Tanbir Chowdhury** under the supervision of **Md Rafid Haque**, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Md Rafid Haque
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Tamim Ahmed
Student ID: 200041150
Date: September 29, 2025

Kazi Samin Yasar Alam
Student ID: 200041119
Date: September 29, 2025

Md Tanbir Chowdhury
Student ID: 200041114
Date: September 29, 2025

This page has been intentionally left blank

Contents

Abstract

This thesis focuses on detecting humor, sarcasm, offensiveness, and vulgarity in Bangla-English code-mixed text, an area largely overlooked in existing natural language processing (NLP) research. A novel dataset has been proposed, which will be created by scraping and filtering social media content, followed by manual annotation across four attributes. Two transformer-based approaches were explored in small scale: multi-class and multi-label text classification. The study also proposes future directions, including dataset balancing, comparative evaluation of transformer models and large language models (LLMs), and the introduction of a *SarOff Score* to better capture sarcasm-offense overlap. By addressing the complexities of code-mixed tone detection, this work advances NLP in low-resource, multilingual settings.

Chapter 1

Introduction

In today's digital world, multilingual communication is no longer the exception but the norm, especially across social media platforms where users often blend languages seamlessly. In South Asia, Bangla-English code-mixing has emerged as a dominant form of informal written expression, reflecting deep cultural integration and fluid linguistic behavior. However, natural language processing (NLP) technologies have struggled to keep pace with this phenomenon. While significant advances have been made for monolingual English or resource-rich languages, code-mixed texts remain a frontier marked by transliteration errors, informal spellings, ambiguous language cues, and a rich mixture of cultural references. These challenges become even more acute when the task involves detecting complex, subjective phenomena like humor, sarcasm, offense, and vulgarity — all of which are heavily dependent on cultural nuance and linguistic interplay.

This thesis proposes MixSarc, the first large-scale, annotated Bangla-English code-mixed corpus specifically designed for humor, sarcasm, offense, and vulgarity detection. Beyond simply building a dataset, the work establishes strong baselines across these tasks using both traditional and transformer-based models. It further explores how understanding sarcasm and humor can substantially improve sentiment analysis models in code-mixed environments, and how offense and vulgarity are expressed differently depending on sarcastic intent. By focusing on scalable, culturally aware modeling approaches, this research aims to fill a critical gap in the study of South Asian code-mixed language and advance NLP capabilities for increasingly multilingual digital landscapes.

1.1 Objectives

The primary objective of this thesis is to construct a comprehensive, high-quality annotated dataset — MixSarc — for the Bangla-English code-mixed environment, focused on humor, sarcasm, offense, and vulgarity detection. Existing resources for these tasks are either scarce, synthetic, or narrowly domain-specific, preventing robust model development. MixSarc aims to provide an organic, balanced dataset that better captures the diversity and complexity of real-world social media language, serving as a new benchmark for future research in this area.

In addition to corpus creation, the thesis seeks to establish strong baseline results using a range of modeling techniques, from classical machine learning models to modern transformer architectures. The goal is to measure how different approaches perform on this challenging new dataset and to lay a foundation for further improvements. Special emphasis will be placed on how sarcasm and humor understanding can enhance sentiment analysis models — a crucial direction, as sarcasm often flips the intended polarity of statements, leading to sentiment classification errors if not properly detected.

Finally, the work aims to deeply analyze how offense and vulgarity are expressed within sarcastic and humorous contexts. Understanding the relationship between humor and offense is vital not only for better classification performance but also for developing culturally sensitive, ethical NLP systems. This thesis aspires to set a new standard for code-mixed resource building and model evaluation in Bangla-English contexts, while offering practical insights into the linguistic interplay of humor, sentiment, and social norms.

1.2 Motivation

Bangla-English code-mixing is a ubiquitous phenomenon across South Asian social media, messaging platforms, and casual digital communication. Yet despite its dominance in everyday interactions, Bangla-English code-mixed text remains critically underrepresented in NLP research. Most state-of-the-art models have been developed primarily for monolingual English or other high-resource languages, leaving a substantial gap in their ability to handle the transliteration inconsistencies, slang, ambiguous vocabulary, and cultural blending inherent to code-mixed data. Without dedicated resources and modeling efforts, NLP systems fail to generalize effectively in these environments.

Adding to the complexity, humor, sarcasm, and offense are deeply subjective constructs that vary widely across cultures, age groups, and social contexts. Unlike objective tasks such as named entity recognition or part-of-speech tagging, detecting humor or sarcasm requires a nuanced understanding of tone, context, and cultural references. These challenges are amplified in code-mixed settings where language choice itself can signal irony, social stance, or emotional intent. Yet current models either ignore these subtleties or treat code-mixed humor as an afterthought, resulting in poor real-world applicability.

Motivated by these gaps, this thesis aims to bring focused attention to sarcasm, humor, offense, and vulgarity detection in Bangla-English code-mixed text. The work seeks to adapt and evaluate transformer-based models specifically for this low-resource, high-complexity scenario, recognizing that success here will not only advance technical benchmarks but also pave the way for more culturally aware, socially responsible NLP applications.

1.3 Problem Statement

Despite the widespread use of Bangla-English code-mixing in online discourse, existing NLP models struggle to accurately process such hybrid language, especially when dealing with subjective phenomena like sarcasm, humor, offense, and vulgarity. The lack of large, high-quality annotated datasets for these tasks severely hampers the development of robust models. Furthermore, most previous research has either relied on synthetic data, monolingual assumptions, or adaptations of multilingual models not specifically tuned to the nuances of code-mixed expression.

Current sentiment analysis systems often misinterpret sarcastic or humorous posts, leading to major errors in sentiment polarity detection. Similarly, detecting offense and vulgarity becomes significantly harder when sarcasm masks the true intent of the text. Without systematic benchmarking and dataset-driven analysis, it is impossible to build models that handle these nuances reliably. This thesis addresses these gaps by creating MixSarc, a large annotated resource for Bangla-English humor and sarcasm detection, and by evaluating scalable modeling strategies that are better suited to the cultural and linguistic complexity of real-world code-mixed communication.

Chapter 2

Related Works

This chapter provides a comprehensive overview of the existing literature relevant to the identification of humor, sarcasm, offense, and vulgarity in Bangla-English code-mixed text. As social media increasingly becomes a space for dynamic, multilingual communication, understanding the intricacies of code-mixing and the linguistic nuances that accompany subjective expressions like humor and offense is critical for advancing natural language processing (NLP) systems.

The review is structured into four key sections to systematically contextualize the present research. First, we examine prior studies on code-mixing in Bangla-English, exploring the linguistic properties of code-mixed communication and early efforts in language identification and normalization. Second, we discuss sentiment analysis in code-mixed texts, highlighting how mixed-language input complicates polarity detection and how earlier models attempted to adapt to such noisy environments. Third, we focus on sarcasm and humor detection, tracing the evolution from rule-based techniques to deep learning and transformer-based approaches, while emphasizing the scarcity of code-mixed resources. Finally, we review literature on offensiveness and vulgarity detection, particularly in the context of sarcasm, and discuss how nuanced interpretation is necessary when intent is masked by humor.

By synthesizing insights from these diverse but interconnected domains, this chapter not only positions the current research within the broader landscape of NLP but also critically identifies the gaps — particularly the lack of large-scale annotated datasets and culturally sensitive modeling — that this thesis seeks to address.

2.1 Code Mixing in Bangla-English

Code-mixing refers to the fluid switching between two or more languages within a single discourse, often at the level of phrases, sentences, or even individual words. In multilingual societies such as South Asia — and Bangladesh in particular — code-mixing has become an integral part of informal communication, especially in digital contexts like social media, messaging apps, and online forums. In the specific case of Bangla and English, this phenomenon, commonly dubbed "Banglish," has been shaped by a combination of historical colonial influences and contemporary globalized culture. It presents unique challenges to Natural Language Processing (NLP) systems, arising from non-standardized grammar, unconventional spellings, transliterations (e.g., writing Bangla words using the Latin alphabet), and highly context-dependent lexical meanings. Unlike clean monolingual text, code-mixed content often defies conventional linguistic boundaries, making basic tasks like language identification, parsing, and classification far more complex.

Recognizing these challenges early on, Chanda et al. (2016) made one of the pioneering contributions to processing Bangla-English code-mixed data. In their work, Unraveling the English-Bengali Code-Mixing Phenomenon, the authors proposed a novel predictor-corrector model for word-level language identification. The "predictor" phase tentatively assigned language tags to each word using dictionary lookups and statistical n-gram models, while the "corrector" phase refined these predictions by analyzing the surrounding linguistic context to resolve ambiguities, particularly for words that are common to both languages (e.g., "hole" in English vs. "hole" meaning "if" in Bangla). Chanda et al. also created and publicly released a new Facebook chat corpus, which captured naturally occurring code-mixed conversations from Bangladeshi users. This corpus served as a critical resource for evaluating language identification models on real-world, informal data. Their hybrid approach — combining dictionary-based rules, Bengali suffix detection, English n-gram fallback strategies, and supervised machine learning classifiers — achieved 91.65% accuracy using an IBk (k-Nearest Neighbors) classifier, a substantial improvement over earlier approaches such as Das and Gambäck's (2014) system, which achieved only 76.37% accuracy. Chanda et al.'s method demonstrated that integrating rule-based heuristics with data-driven machine learning can substantially improve code-mixed language processing performance, especially when annotated resources are scarce. [6]

As research on code-mixed NLP progressed, attention gradually shifted from rule-based and traditional ML methods to deep learning architectures like LSTM and BiLSTM models. While these models were more adept at capturing sequential dependen-

cies and could learn contextual representations directly from data, they still struggled with the highly noisy, transliterated, and domain-specific nature of Bangla-English code-mix. Furthermore, deep models typically required large quantities of labeled data, which was a major bottleneck in this low-resource setting.

In response to these limitations, Raihan et al. (2024) introduced a significant advancement with their work on Mixed-Distil-BERT: Code-mixed Language Modeling for Bangla, English, and Hindi. Recognizing the lack of effective pretrained models for code-mixed environments, Raihan et al. developed two compact yet powerful models: Tri-Distil-BERT and Mixed-Distil-BERT.[9]

Tri-Distil-BERT was pretrained on standard monolingual corpora in English, Bangla, and Hindi, aiming to learn strong multilingual representations in a compact architecture.

Mixed-Distil-BERT went a step further by introducing a second stage of pretraining on synthetically generated English–Bangla–Hindi code-mixed text, simulating the natural patterns of language mixing found in social media conversations.

To facilitate this, Raihan et al. also created new code-mixed datasets, each containing approximately 100,000 samples, specifically annotated for emotion detection, sentiment analysis, and offensive-language identification. This resource contribution was crucial, addressing the chronic data scarcity problem in code-mixed research.

Their two-stage pretraining strategy — first on monolingual data, then on synthetic code-mixed text — proved highly effective. Despite being a compressed model (based on DistilBERT, which has roughly 40% fewer parameters than BERT), Mixed-Distil-BERT achieved competitive or even superior results compared to much larger multilingual models like mBERT and XLM-R, across multiple downstream tasks. This work demonstrated that task-specific code-mixed pretraining — even on synthetic data — could significantly boost performance, offering a lightweight and practical solution for real-world multilingual NLP applications. It also highlighted that code-mixing is not merely a "noise problem" to be handled heuristically but rather a linguistic reality that demands dedicated modeling strategies.

Together, these works — Chanda et al.'s hybrid predictor-corrector approach and Raihan et al.'s compact transformer-based modeling — mark important milestones in the evolving landscape of Bangla-English code-mixed NLP. They illustrate the transition from manually crafted rule-based systems to sophisticated neural architectures specifically adapted to the complex, dynamic realities of South Asian digital communication.

2.2 Sentiment Analysis in Bangla-English Code Mixed Language

Sentiment analysis in code-mixed languages has become an increasingly important research area, given its critical applications in social media monitoring, customer feedback analysis, and broader opinion mining. The unique linguistic diversity and complex interplay between languages in code-mixed texts necessitate specialized computational approaches that can handle transliterations, spelling variations, and context-dependent semantics. While significant progress has been made for languages like English–Hindi or English–Spanish, Bangla-English code-mixed sentiment analysis has historically suffered from inadequate attention, primarily due to a lack of high-quality datasets and dedicated modeling efforts.

One of the early efforts to address multilingual code-mixed sentiment analysis was SentMix-3L, introduced by Raihan et al. (2023). SentMix-3L represented a pioneering tri-lingual dataset designed for handling Bangla-English-Hindi code-mixed texts — an environment that mirrors the linguistic reality of South Asian social media far more accurately than traditional bilingual datasets. Unlike earlier resources that often limited themselves to two languages, SentMix-3L captured utterances blending three languages within a single discourse. The dataset consisted of 1,007 naturally generated posts, constructed by multilingual contributors who mixed Bangla, English, and Hindi phrases in authentic, informal ways. Each post was manually annotated with a sentiment label — positive, neutral, or negative — through a rigorous two-stage process to ensure high annotation quality, achieving strong inter-annotator agreement. [10]

Recognizing the small size of the natural dataset, the authors also developed a synthetic corpus of 100,000 samples, generated by applying code-mixing algorithms to the Amazon Reviews corpus. These synthetic examples served as training data, while the natural SentMix-3L samples were used for evaluation. Experimental results revealed that zero-shot prompting with GPT-3.5 outperformed traditional transformer models, achieving a weighted F1-score of 0.62, while XLM-R achieved 0.59, and fine-tuned models like BanglishBERT and mBERT scored around 0.56. These findings highlighted both the promise of large language models and the difficulty posed by authentic tri-lingual code-mixed sentiment tasks. SentMix-3L marked an important step forward, offering a realistic and challenging benchmark for future research, but its reliance on synthetic data and limited natural samples left open important gaps in capturing the true diversity of real-world code-mixing.

To address these shortcomings, BnSentMix was introduced through the paper "BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis". Recognizing that earlier datasets, including SentMix-3L, suffered from limited scale, synthetic construction, or restricted domain coverage, the creators of BnSentMix aimed to offer a much larger, naturally occurring, and informal-domain corpus for Bangla-English code-mixed sentiment analysis. [4]

The final BnSentMix dataset contained 20,000 instances carefully collected from real-world sources including Facebook, YouTube, and e-commerce product reviews. A rigorous scraping, cleaning, and filtering pipeline was implemented to ensure that only genuine code-mixed texts were retained. A fine-tuned multilingual BERT-based code-mix detection model — achieving 94.6% accuracy — was used to automatically identify valid Bangla-English code-mixed samples, followed by manual verification and annotation by native Bengali speakers. Each instance was labeled as either positive, negative, neutral, or — uniquely — mixed, the latter category capturing utterances that expressed both positive and negative sentiments simultaneously. High annotation consistency was ensured through a Cohen's score of 0.86, demonstrating strong inter-annotator agreement.

To benchmark the dataset, a wide range of models were evaluated, from traditional machine learning classifiers (such as Logistic Regression and SVMs) to neural networks (RNNs, LSTMs) and modern transformer-based models (BERT, mBERT, XLM-R, BanglaBERT, and BanglishBERT). The best performance was achieved by a BERT model further pre-trained on the BnSentMix corpus (BERT-CMB), which reached approximately 69.8% accuracy. This highlighted not only the difficulty of sentiment analysis in noisy, informal code-mixed environments but also the importance of domain-specific pretraining for boosting performance.

Compared to previous works, BnSentMix made several important contributions: it offered a significantly larger dataset, introduced the novel "mixed" sentiment class, ensured high-quality manual annotations, and was made publicly available for the wider research community. By combining natural language diversity with robust benchmarking, BnSentMix set a new standard for code-mixed sentiment analysis resources in low-resource languages like Bangla. It filled key gaps left by SentMix-3L, particularly in offering a realistic, informal, and naturally constructed dataset suitable for developing scalable, real-world models.

Despite these important advances, both SentMix-3L and BnSentMix reflect a broader trend: while interest in code-mixed language processing is growing, the collection of high-quality, annotated datasets remains a significant bottleneck, especially for

low-resource languages like Bangla. Furthermore, while sentiment analysis has received increasing attention, tasks such as humor and sarcasm detection — especially in the context of Bangla-English code-mixed text — remain largely underexplored. This underscores a major gap that the present thesis seeks to fill: constructing a large, high-quality annotated dataset specifically tailored to humor, sarcasm, offense, and vulgarity detection in Bangla-English code-mixed environments, thereby pushing the frontiers of multilingual and multimodal NLP research.

2.3 Sarcasm and Humor Detection in Code Mixed Language (Bangla-English)

Sarcasm detection stands as one of the most challenging tasks in Natural Language Processing (NLP), primarily due to the complex interplay of context, speaker intent, cultural background, and the frequent reliance on non-literal language. Unlike traditional sentiment analysis, which often categorizes text based on surface-level cues, sarcasm requires an understanding of the underlying dissonance between what is said and what is meant. For example, a sentence like "You are very healthy, I think you can go on for months without eating" would likely be misclassified by a naïve sentiment analysis model as expressing positive sentiment, despite its clearly sarcastic undertone. The difficulty amplifies in code-mixed languages, where speakers fluidly combine words and structures from multiple languages within the same sentence, introducing further linguistic and cultural nuances. Consider, for example, the Bengali-English code-mixed sentence: "Tor moto sharadin porashona korle, ami ajke Harvard e first hoitam," — a sarcastic remark implying mockery rather than genuine praise, which would likely be misunderstood without cultural and contextual awareness.

2.3.1 Early Foundations: Sarcastic Incongruity and Pattern Bootstrapping

One of the pioneering works that formalized sarcasm detection as a structured NLP task was that of Riloff et al. (2013). They proposed a pattern-bootstrapping approach for identifying sarcasm in Twitter posts. Grounded in the theory of sarcastic incongruity, their approach focused on detecting contradictions between the expressed positive sentiment and an underlying negative situation. Their method started with seed lists of overtly positive sentiment words, such as "great" and "wonderful," and iteratively expanded these by mining linguistic patterns that suggested sarcastic intent. For instance, tweets pairing positive adjectives with undesirable situations (e.g., "an-

other Monday morning") were indicative of sarcasm. A key innovation of Riloff et al.'s work was the automated pattern expansion, which minimized manual annotation efforts and allowed the model to adapt to emerging sarcastic expressions over time. Their contribution laid an essential theoretical and methodological groundwork for future sarcasm detection models, emphasizing that the crux of sarcasm often lies not in lexical cues alone, but in the contrast between sentiment and context [11]

2.3.2 Introducing Code-Mixed Sarcasm: English-Hindi Developments

Building upon the insights from monolingual sarcasm detection, Swami et al. (2018) shifted focus to the unique challenges posed by code-mixed data, particularly English-Hindi social media content. Recognizing a critical resource gap, they created one of the first manually annotated corpora specifically designed for sarcasm detection in code-mixed tweets. Their dataset comprised approximately 5,000 English-Hindi tweets, meticulously filtered to ensure both code-mixing and sarcastic intent through indicators like hashtags (sarcasm, irony). Importantly, the annotation process involved manual validation, thus enhancing the reliability of the corpus compared to earlier hashtag-based methods. To establish performance benchmarks, Swami et al. evaluated a variety of machine learning classifiers, including Support Vector Machines (SVM) and Random Forests, employing feature sets like character n-grams and word-level embeddings. Achieving a best accuracy of 78 4%, their work demonstrated the feasibility of sarcasm detection in code-mixed environments and paved the way for more sophisticated bilingual and multilingual sarcasm models. [13]

2.3.3 Scaling Up and Deep Learning: Aggarwal et al. (2020)

Taking this endeavor further, Aggarwal et al. (2020) recognized that effective sarcasm detection in code-mixed settings required not just small datasets, but large-scale, linguistically rich corpora and advanced deep learning techniques. They developed a substantial dataset containing approximately 107,000 Hinglish tweets, balanced evenly between sarcastic and non-sarcastic instances. Crucially, they introduced bilingual word embeddings trained on a mix of standard English and code-mixed Hinglish texts. This allowed their models to better capture the semantic richness and idiosyncrasies of code-mixed language. They experimented with multiple deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), Bi-directional LSTMs (Bi-LSTMs), and attention-enhanced Bi-LSTMs. The best performing model — an attention-based Bi-LSTM —

achieved an impressive 78.5% accuracy, outperforming traditional machine learning baselines. Notably, their results revealed that whole-word context modeling (Word2Vec) was more effective than subword modeling (FastText) in code-mixed scenarios, emphasizing the need for embeddings that preserve broader semantic context. Beyond just empirical results, Aggarwal et al. contributed valuable public datasets and resources, pushing the boundaries of multilingual and code-mixed NLP research. [2]

2.3.4 Capturing True Intent: The iSarcasm Dataset

While previous sarcasm datasets often relied heavily on user hashtags as proxy indicators, Oprea and Magdy (2020) challenged this paradigm with the introduction of the iSarcasm dataset. They argued that many tweets labeled sarcastic based on hashtags did not necessarily express intended sarcasm — a significant limitation for training reliable models. To rectify this, they adopted a crowdsourced annotation approach, meticulously selecting tweets where sarcasm was truly intended by the author. Their resulting dataset, although smaller (about 4,000 tweets), was considerably cleaner and more focused. Experiments using Bi-LSTMs and transformer-based models like BERT revealed that performance dropped when models were tested against this more rigorously annotated dataset compared to noisy, hashtag-based datasets. This highlighted an important insight: genuine sarcasm detection requires deeper contextual and pragmatic understanding, not mere reliance on superficial cues. The iSarcasm dataset thus served as a new gold standard for studying intended sarcasm[8]

2.3.5 Bengali Language Initiatives: Ben-Sarc and BanglaSarc

Parallel to English-Hindi efforts, researchers also began addressing sarcasm detection in Bengali, another major South Asian language. Chowdhury et al. (2020) introduced Ben-Sarc, an early Bengali sarcasm corpus created through self-annotation, where sarcastic comments were identified based on user-generated hashtags and linguistic hints. While this approach enabled the rapid construction of a reasonably large dataset, it inevitably introduced noise — some tweets labeled sarcastic might not have genuinely conveyed sarcasm. Nonetheless, Ben-Sarc provided a critical starting point for research in a largely unexplored language space. Baseline experiments showed that deep learning models like LSTMs and CNNs performed better than traditional classifiers, reinforcing global trends seen in sarcasm detection.[7]

To address Ben-Sarc's noisiness, Hasan et al. (2021) curated BanglaSarc, a manually annotated Bengali sarcasm dataset with about 3,500 posts. Their human-centric annotation process ensured higher data quality, making BanglaSarc a more reliable

benchmark. They evaluated traditional machine learning models like Logistic Regression, SVMs, and Random Forests using feature representations such as TF-IDF and word embeddings. Although the models achieved only moderate accuracy, BanglaSarc established a crucial foundation for building more sophisticated sarcasm detection models in Bengali. Together, Ben-Sarc and BanglaSarc demonstrated two complementary approaches — one prioritizing scale and another prioritizing quality — thereby advancing sarcasm research in Bengali significantly.[5]

2.3.6 Humor and Offense Detection: The DuluthNLP System

Expanding beyond pure sarcasm, Akrah (2021) developed the DuluthNLP system for SemEval-2021 Task 7, focusing on detecting humor and offense in social media text. Leveraging the pretrained RoBERTa transformer model, the system underwent extensive hyperparameter tuning, initially heuristic but later enhanced through Bayesian optimization. This systematic hyperparameter search led to significant performance improvements, with humor detection F1-scores rising from 0.939 to 0.957. Akrah's work highlighted that fine-tuning powerful pretrained models for specific subtasks — along with careful hyperparameter optimization — is critical to achieving robustness in detecting humor and offense, both of which share fuzzy boundaries with sarcasm. [3]

2.3.7 Data Augmentation for Sarcasm: The UTNLP System

Finally, Abaskohi et al. (2022) presented the UTNLP system for SemEval-2022 Task 6, offering new insights into data augmentation strategies for sarcasm detection. They compared two approaches: generative augmentation using GPT-2 to synthesize new sarcastic/non-sarcastic tweets, and mutation-based augmentation involving operations like word shuffling, deletion, and synonym replacement. Testing models from SVMs to RoBERTa and Google's T5, they found that mutation-based methods consistently outperformed generative ones. Their best model—a RoBERTa fine-tuned with mutation-augmented data—achieved an F1-sarcastic score of 0.414. This study illuminated the delicate balance between generating realistic sarcastic content and preserving critical sarcastic cues, reaffirming the effectiveness of controlled, mutation-based augmentations combined with transformer architectures. [1]

Despite notable advancements, Bangla-English code-mixed sarcasm detection remains an underexplored territory. Lessons from Hindi-English studies—such as the benefits of attention mechanisms, bilingual embeddings, and high-quality annotations—hold promising potential for Bangla-English research. However, the lack of a dedicated

Bangla-English sarcasm dataset marks a crucial gap that this thesis seeks to fill, contributing to the broader goal of making sarcasm detection truly multilingual and contextually aware.

2.4 Offensiveness and Vulgarity Detection

The detection of offensive and vulgar language in Bangla-English code-mixed content is a relatively young but rapidly maturing research area, driven by the proliferation of social media and the urgent need to moderate harmful online discourse. Code-mixing introduces unique challenges such as transliteration inconsistencies, spelling variations, and complex syntactic blending between two languages, making traditional monolingual approaches insufficient.

One of the pioneering efforts in this field was led by Mandal et al. (2018) as part of the ICON-2018 Shared Task on Sentiment Analysis for Indian Languages. Recognizing the scarcity of annotated resources for code-mixed offensive content, the authors created a Bangla-English corpus annotated at both the word and sentence levels for offensive language markers. Their experiments employed a variety of traditional machine learning models, notably Support Vector Machines (SVM), Random Forests, and ensemble methods. The ensemble approaches, combining predictions from multiple classifiers, showed particularly promising results by leveraging the strengths of different algorithms. A key observation from their study was that simple lexical features like TF-IDF vectors, n-grams, and handcrafted syntactic features still held considerable predictive power even in code-mixed contexts. However, they also emphasized that the lack of standardized spelling and grammar in informal social media communication posed significant barriers to achieving high performance, hinting at the need for more robust feature engineering or data-driven modeling.

Building upon these early insights, Sazzed (2020) took a more specialized and large-scale approach by developing the Bengali-English Offensive Language (BEOL) dataset. This resource marked a significant advancement in the field by offering a substantial corpus of over 10,000 Facebook comments annotated into three categories: offensive, non-offensive, and vulgar. The meticulous annotation process paid special attention to differentiating vulgarity from general offensiveness, acknowledging that vulgar language often has a distinct stylistic and social impact. Sazzed established strong baseline performances using deep learning architectures, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, both well-suited to capturing sequential dependencies in noisy, informal text. His experiments

demonstrated that deep neural models outperformed traditional machine learning techniques, especially when fed with word embeddings pretrained on similar code-mixed corpora. Importantly, the study showed that the incorporation of transliteration handling, stop-word normalization across languages, and custom tokenization significantly boosted model accuracy. Sazzed's work thus provided both a critical data resource and a methodological blueprint for future offensiveness detection in Bangla-English settings.

As the field evolved, researchers increasingly shifted towards leveraging the capabilities of transformer-based architectures, which had already revolutionized many other NLP tasks. Hasan et al. (2021) made notable contributions by fine-tuning multilingual BERT (mBERT) models specifically on code-mixed Bangla-English data for abusive language detection. Unlike previous works that primarily relied on handcrafted features or shallow neural networks, Hasan et al. embraced the contextual richness offered by transformers, which can capture complex interdependencies between words regardless of code-switching or transliteration variations. Their methodology involved careful preprocessing steps such as script unification (converting Romanized Bangla words back to native script when possible), noise removal, and balanced resampling to address class imbalance—a common problem where non-offensive instances greatly outnumber offensive ones. By fine-tuning mBERT, they achieved significant performance improvements over both traditional baselines and earlier deep learning approaches, illustrating that pretrained multilingual models can adapt surprisingly well to noisy, low-resource code-mixed domains when appropriately fine-tuned. Another important insight from their study was the effectiveness of domain-specific pretraining: additional masked language modeling on code-mixed corpora before fine-tuning the classification head led to measurable gains.

Together, these successive efforts — from traditional classifiers to deep recurrent networks and ultimately to transformer-based models — highlight several key lessons for offensive language detection in Bangla-English code-mixed environments:

Specialized preprocessing is crucial to address informal writing styles, spelling inconsistencies, and transliteration issues.

Large, well-annotated datasets like BEOL serve as indispensable resources for both benchmarking and model training.

Deep learning architectures, especially transformers fine-tuned on domain-specific data, now represent the cutting edge, offering significant improvements in understanding nuanced, multilingual expressions of offense and vulgarity.

Despite this progress, challenges remain. Sarcasm intertwined with offensive language, cultural subtleties in vulgar expressions, and the dynamic evolution of slang terms are areas where current models still struggle. Future research will likely focus on augmenting existing corpora with richer pragmatic annotations, designing pretraining objectives better suited to code-mixed inputs, and exploring few-shot or zero-shot learning approaches to quickly adapt to new linguistic phenomena without requiring massive labeled datasets.

2.5 Identify Gaps and Opportunities

Despite significant advancements in the fields of code-mixed language processing, sentiment analysis, sarcasm detection, humor recognition, and offensiveness detection, the current body of literature reveals several critical gaps, particularly when it comes to Bangla-English code-mixed text. Although research on Hindi-English code-mixing has gained momentum over the last few years, the Bangla-English scenario remains relatively underexplored, leaving numerous opportunities for impactful contributions.

One of the most prominent gaps is the lack of specialized datasets tailored for Bangla-English code-mixed content beyond basic sentiment analysis. While resources such as BnSentMix and SentMix-3L have provided important starting points for sentiment classification, there remains an absence of datasets that specifically target humor, sarcasm, offensiveness, and vulgarity in a Bangla-English code-mixed context. Sentiment is only one dimension of human communication; without accounting for complex tones such as sarcasm or varying degrees of offensiveness, models remain shallow and incomplete in their understanding.

Even in existing datasets, major issues have been identified. To illustrate this, a small-scale experiment was conducted on BnSentMix, regarded as the most robust Bangla-English code-mixed sentiment dataset currently available. Two stark issues emerged:

The model exhibited a heavy bias toward classifying sarcastic sentences as negative sentiment. In fact, approximately 51

Pretrained models based on BnSentMix struggled immensely when exposed to offensive or vulgar content. Instead of correctly identifying such expressions as predominantly negative, the models confusedly distributed the labels among positive, negative, neutral, and mixed categories with almost equal probability, reflecting poor real-world robustness.

The following table summarizes the classification distribution results when the BnSentMix models were tested on a smaller version of the intended dataset:

Table 2.1: Performance of BnSentMix pretrained models on offensive and vulgar Bangla-English code-mixed content.

Type	Positive	Negative	Neutral	Mixed	Total
Both Offensive and Vulgar	2	6	1	0	9
Just Offensive	21	32	12	4	69
Just Vulgar	13	25	16	16	70

The table clearly reveals that the existing systems lack robustness when tasked with decoding harsher tones, mixing up vulgarity and offensiveness with positive or neutral sentiments at an alarming rate.

Another major gap lies in the focus on other language pairs, particularly Hindi-English. The overwhelming majority of code-mixed research to date has concentrated on Hindi-English, likely due to the larger Hindi-speaking population and greater availability of data. However, models trained on Hindi-English code-mixed text do not seamlessly transfer to Bangla-English because of differences in syntax, vocabulary, culture, and code-switching patterns. The nuances of Bangla-English code-mixing—such as transliteration peculiarities, morphology, and the social use of humor and sarcasm—remain insufficiently understood.

Offensiveness research, especially, remains underdeveloped for Bangla-English settings. While isolated efforts have been made for Bangla-only content, and some preliminary work has been conducted on Bangla-English offensive language detection, virtually no studies systematically address how sarcasm interacts with offensive language in a code-mixed environment. This is critical because sarcastic remarks often mask or amplify offensiveness in subtle ways, making detection much harder if models rely only on surface-level word cues.

Furthermore, a persistent monolingual bias continues to shape available resources. Datasets like Ben-Sarc (for Bengali sarcasm) and various English-only sarcasm corpora fail to represent the messy, rich reality of code-mixed discourse where users fluidly alternate between Bangla and English even within a single sentence. Such monolingual corpora provide limited utility when training models to handle bilingual or hybrid expressions.

In light of these gaps, this thesis proposes to make the following novel contributions:

The creation of a Bangla-English code-mixed dataset specifically annotated for humor, sarcasm, and graded levels of offensiveness and vulgarity.

The development of baseline models and deep learning architectures trained and evaluated on this dataset to establish benchmark performances.

A critical analysis of how traditional sentiment models behave under the stress of nuanced, offensive, or sarcastic code-mixed inputs, thereby identifying failure modes and areas for improvement.

The exploration of attention mechanisms, bilingual embeddings, and transformer-based approaches to better model the contextual and pragmatic subtleties inherent in code-mixed communication.

By addressing these research gaps, the thesis aims not only to enrich the available resources for Bangla-English code-mixed NLP but also to push the field forward toward more nuanced, culturally aware, and context-sensitive natural language understanding systems.

2.6 Summary

This literature review synthesizes a broad range of research related to humor, sarcasm, sentiment, and offensiveness detection in code-mixed Bangla-English text, offering a comprehensive understanding of current advancements and identifying persisting gaps in natural language processing (NLP). By systematically covering four major focus areas — code-mixing phenomena, sentiment analysis, sarcasm and humor detection, and offensiveness and vulgarity detection — the review establishes a clear context for the contributions of this thesis.

In the area of code-mixing, early work by Chanda et al. (2016) laid the groundwork by developing robust word-level language identification models using a predictor-corrector framework, combining rule-based techniques and machine learning. Their introduction of a Facebook chat corpus marked a significant step in building resources for handling Bangla-English mixed text. More recent research by Raihan et al. (2023, 2024) pushed the frontier further, addressing tri-lingual code-mixed scenarios involving Bangla, English, and Hindi. Their innovations — particularly the development of Mixed-DistilBERT models and synthetic code-mixed datasets — demonstrated the effectiveness of two-stage pretraining (multilingual first, then code-mixed) for achieving state-of-the-art performance even with relatively lightweight transformer architectures. These works collectively underscore the importance of specialized pretraining for multilingual, informal, and code-mixed contexts.

In the domain of sentiment analysis, studies such as SentMix-3L introduced by Raihan

et al. (2023) and BnSentMix have been instrumental in creating valuable benchmarks for code-mixed sentiment detection. SentMix-3L notably captured the linguistic richness of South Asian digital communication through tri-lingual code-mixed examples. However, its heavy reliance on synthetic data limits its representational authenticity. In contrast, BnSentMix offered a larger, naturally sourced Bangla-English dataset, achieving high inter-annotator agreement and incorporating a unique "mixed" sentiment label to reflect the nuanced emotions often conveyed in social media text. Despite these advancements, small experiments on BnSentMix revealed that pre-trained models exhibit significant biases — misclassifying sarcastic expressions predominantly as negative and demonstrating confusion when dealing with offensive or vulgar content, often distributing incorrect labels (positive, negative, neutral, mixed) almost equally.

Sarcasm and humor detection, an inherently challenging subfield due to its reliance on cultural knowledge and pragmatic understanding, has seen encouraging progress in Hindi-English contexts but remains severely underexplored for Bangla-English. Pioneering work by Riloff et al. (2013) introduced the notion of sarcastic incongruity, inspiring methods that detect the mismatch between surface-level sentiment and contextual reality. Building on this, Swami et al. (2018) and Aggarwal et al. (2020) contributed significantly by creating code-mixed Hindi-English sarcasm datasets and leveraging bilingual embeddings with deep learning models like BiLSTM-Attention to achieve notable improvements. However, analogous resources for Bangla-English remain absent. Bengali monolingual datasets like Ben-Sarc (Chowdhury et al., 2020) and BanglaSarc (Hasan et al., 2021) made initial strides, but their focus on pure Bengali texts does not fully capture the nuances of code-mixed communication prevalent on Bangladeshi social media.

The area of offensiveness and vulgarity detection similarly reveals a promising yet incomplete trajectory. Early contributions such as Mandal et al. (2018) initiated exploration into Bangla-English offensive content using traditional machine learning classifiers. Later, Sazzed (2020) created the Bengali-English Offensive Language (BEOL) dataset, significantly expanding the scale of available resources and establishing baselines using LSTM and GRU models. More recently, Hasan et al. (2021) demonstrated that fine-tuning multilingual transformers like mBERT can yield substantial improvements for abusive language detection in code-mixed settings. However, the literature largely neglects deeper issues — particularly the detection of offensiveness within sarcastic or humorous contexts, where offensiveness may be masked or intensified through non-literal expressions.

Across all these domains, several critical gaps consistently emerge:

There is a clear lack of specialized Bangla-English datasets targeting humor, sarcasm, offensiveness, and vulgarity in code-mixed social media texts.

Most sarcasm research remains focused on Hindi-English, often assuming cultural patterns and code-mixing behaviors that do not align with Bangla-English dynamics.

Existing sentiment models, including those trained on BnSentMix, struggle with nuanced tone detection, showing substantial biases when sarcasm, vulgarity, or offensiveness are present.

Offensiveness research, although improving, remains largely monolingual and rarely integrates the contextual richness of humor and sarcasm, leading to models that miss subtleties in human communication.

Finally, there remains a general monolingual bias in available resources, with Bengali and English considered separately, failing to reflect real-world code-mixed discourse where language blending is fluid, frequent, and culturally specific.

Recognizing these gaps, the current thesis aims to address these deficiencies by developing a novel Bangla-English code-mixed dataset annotated explicitly for humor, sarcasm, and graded levels of offensiveness and vulgarity. Through careful design, annotation, and model benchmarking, this work seeks not only to advance the quality of resources available for Bangla-English code-mixed NLP but also to push the boundary of nuanced tone and sentiment detection in multilingual, informal, and highly context-dependent environments.

This contribution is expected to bridge the current disconnect between existing research and real-world needs, enabling future NLP systems to better understand, moderate, and interact within Bangla-English code-mixed digital spaces — ultimately supporting the broader goal of building inclusive, culturally aware, and linguistically robust language technologies.

Table 2.2: Summary of Key Literature Reviewed on Code-Mixed Language Processing, Sentiment, Sarcasm, Humor, and Offensiveness Detection-1

Paper	Focus Area	Language(s)	Dataset/Model Contribution	Key Findings
Chanda et al. (2016)	Code-Mixing Detection	Bangla-English	Facebook Chat Corpus; Predictor-Corrector Model	Machine learning combined with rule-based features achieved high word-level code identification accuracy (91.65%).
Raihan et al. (2024)	Code-Mixed Language Modeling	Bangla-English-Hindi	Mixed-DistilBERT and Tri-DistilBERT; Synthetic datasets	Two-stage pretraining improves performance on tri-lingual code-mixed tasks using lightweight models.
Mandal and Das (2018)	Sentiment Analysis	Bangla-English (Romanized)	Movie Reviews Dataset	Code-mixed features significantly improved sentiment classification (SVM 72.5% accuracy).
Raihan et al. (2023)	Sentiment Analysis	Bangla-English-Hindi	SentMix-3L Dataset (natural + synthetic)	Highlighted tri-lingual complexities; GPT-3.5 outperforming transformers on zero-shot prompting.
BnSentMix (2023)	Sentiment Analysis	Bangla-English	BnSentMix Dataset (20K samples)	Introduced "mixed" sentiment label; BERT-CMB model achieved 69.8% accuracy; highlighted real-world code-mix challenges.
Riloff et al. (2013)	Sarcasm Detection	English	Pattern Bootstrapping for Sarcasm	Introduced theory of sarcastic incongruity; contrast between positive sentiment and negative situations.
Swami et al. (2018)	Sarcasm Detection	English-Hindi	Code-Mixed Sarcasm Corpus (5K tweets)	Character n-grams and word features; achieved 78.4% accuracy with traditional ML models.
Aggarwal et al. (2020)	Sarcasm Detection	English-Hindi	Large Hinglish Dataset (107K); Bilingual Embeddings	BiLSTM-Attention model achieved 78.5%; Word2Vec embeddings outperformed FastText.
Oprea and Magdy (2020)	Sarcasm Detection	English	iSarcasm Dataset (Intended Sarcasm)	Models struggled with intended sarcasm; need for deeper pragmatic understanding beyond hashtags.
Chowdhury et al. (2020)	Sarcasm Detection	Bengali	Ben-Sarc (Self-Annotated Corpus)	Demonstrated feasibility of sarcasm detection in Bengali despite noisy annotations.
Hasan et al. (2021)	Sarcasm Detection	Bengali	BanglaSarc (Manually Annotated Corpus)	Cleaner sarcasm dataset; enabled robust baseline evaluations with traditional ML models.
Akrah (2021)	Humor and Offense Detection	English	DuluthNLP for SemEval-2021 Task 7	Fine-tuning RoBERTa + Bayesian optimization improved F1 humor detection score from 0.939 to 0.957.

Table 2.3: Summary of Key Literature Reviewed on Code-Mixed Language Processing, Sentiment, Sarcasm, Humor, and Offensiveness Detection-2

Paper	**Focus Area**	**Language(s)**	**Dataset/Model Contribution**	**Key Findings**
Abaskohi et al. (2022)	Sarcasm Detection with Augmentation	English	UTNLP System (Mutation vs Generative Augmentation)	Mutation-based data augmentation outperformed GPT-2 generation; RoBERTa most effective architecture.
Mandal et al. (2018)	Offensiveness Detection	Bangla-English	ICON-2018 Corpus for Offensive Language	Traditional classifiers showed potential; transliteration and informal text remain challenges.
Sazzed (2020)	Offensiveness and Vulgarity Detection	Bangla-English	BEOL Dataset (10K Facebook Comments)	LSTM and GRU models outperformed traditional baselines; need for special preprocessing identified.
Hasan et al. (2021)	Offensiveness Detection	Bangla-English	mBERT Fine-Tuning on Code-Mixed Data	Fine-tuned transformers significantly outperformed earlier methods; domain-specific adaptation effective.

Chapter 3

MixSarc Dataset

3.1 Data Sourcing

To build a robust Bangla-English code-mixed dataset targeting humor, sarcasm, offensiveness, and vulgarity detection, we curated text data from public Facebook pages known for their humorous and sarcastic content revolving around day-to-day trivial topics. Data scraping was performed using the `Scrapy` web scraping framework.

The following four public Facebook pages were used as sources:

- **Shelby Bhai** – 6,034 data points
- **One Two Three Ami Onek Free** – 8,315 data points
- **Porte Bosh** – 7,386 data points
- **Ammu Dake** – 9,394 data points

A total of 31,129 raw textual data points were collected from these pages.

3.2 Data Cleaning and Preprocessing

Post-scraping, a multi-stage filtering process was implemented to ensure the quality and relevance of the dataset:

1. **Emoji Removal:** All emoticons and emojis were removed using the `replace_emoji` method from Python's `emoji` library.
2. **Script Filtering:** Posts written entirely in Bengali script were removed by detecting Unicode Bengali characters via regular expressions (regex).

3. **Language Identification and Code-Mixed Validation:** A fine-tuned **Multilingual BERT (mBERT)** model was trained on a manually curated dataset distinguishing:

   - Pure English tokenized words
   - Banglish tokenized words (Romanized Bangla)

   Using the mBERT-based classifier, each post was tokenized and filtered based on the following pseudocode:

```
if total_word_count < 4:
    return 0
benglish_percentage = total_benglish_word_count / total_word_count
return 1 if benglish_percentage >= 0.3 else 0
```

   This filtering ensured that at least 30% of the content in a post was Banglish or English, allowing for genuine code-mixed samples to be retained.

3.3 Data Annotation

3.3.1 Annotation Scheme

Following a meticulous filtering process to ensure data quality and relevance, a clean dataset comprising **9,591** data points was curated. Each data point consisted of an individual sentence, carefully selected to meet the criteria for subsequent annotation.

The sentences were manually annotated by human evaluators to assign one or more of the following binary labels, capturing the nuanced characteristics of the content:

- **Humorous**: Indicating the presence of elements intended to provoke amusement or laughter.
- **Sarcastic**: Denoting the use of irony or mockery to convey contempt or ridicule.
- **Offensive**: Signifying content that could reasonably be perceived as insulting, derogatory, or harmful to individuals or groups.
- **Vulgar**: Referring to the inclusion of coarse, profane, or obscene language or themes.

To accommodate the complexity and potential overlap in linguistic expressions, the annotation schema permitted multi-label classifications. For example, a single sen-

tence could be labeled as both humorous and sarcastic, reflecting real-world scenarios where multiple attributes coexist.

To enhance the reliability of the annotations and mitigate individual biases, each sentence was independently evaluated by three distinct annotators. The final label for each data point was determined through a majority voting mechanism, whereby a label was assigned only if at least two of the three annotators agreed on its applicability. This consensus-driven approach ensured consistency and robustness in the labeled dataset, forming a solid foundation for subsequent analysis in this study.

3.3.2 Annotators

Following the initial wishlist and evaluation process, a total of 18 annotators were successfully confirmed. During the evaluation phase, candidates were required to achieve a minimum accuracy of 70% on a sample test to qualify as annotators. The selected annotators represent a diverse age range of 19 to 59 years, ensuring a broad perspective in the annotation task. The group comprises a male-to-female ratio of 11:7, reflecting a balanced gender distribution. These annotators hail from 6 distinct countries and are affiliated with 14 different institutions, further enhancing the diversity and expertise brought to the project. Each annotator were assigned 1,500 sentences to annotate, with payment provided upon successful completion of the task.

3.3.3 Data Statistics

The annotation process, conducted by the confirmed annotators using a majority voting mechanism, resulted in a diverse distribution of labels across the dataset. A significant portion of the 9,591 sentences were identified with single labels, including 750 labeled as humorous, 290 as vulgar, 207 as offensive, and 750 as sarcastic, with an additional 750 sentences marked as having no specific attributes. Multi-label classifications were also prevalent, with 844 sentences annotated as both humorous and sarcastic, the most frequent combination. Other notable combinations included 84 sentences labeled as humorous and vulgar, 76 as humorous and offensive, and 38 as offensive and sarcastic. Less common combinations included 34 sentences marked as vulgar and offensive, 27 as vulgar and sarcastic, 9 as humorous, offensive, and sarcastic, 3 as humorous, vulgar, and sarcastic, and a minimal 2 sentences as vulgar, offensive, and sarcastic. This distribution underscores the complexity of the dataset, reflecting a rich variety of linguistic nuances and overlapping attributes.

3.4 Challenges and Limitations

The annotation process encountered several challenges that impacted the interpretation and classification of the dataset. Notably, sentences labeled as offensive were often found to lack the severity typically associated with such a designation, suggesting a misalignment between the label and the actual content. Similarly, instances classified as vulgar were predominantly characterized by the use of innuendos and slang, rather than overtly crude language. A significant portion of the dataset consisted of taunts and personal jabs, particularly within friend groups, which played a crucial role in elucidating the dynamics of dialogue and colloquial expressions in informal contexts. These findings highlight the difficulty in applying standardized labels to nuanced, context-dependent language, thereby posing limitations to the generalizability of the annotations and necessitating further refinement in future studies to better capture the subtleties of informal communication.

Chapter 4

Methodology

4.1 Methodology

4.1.1 Approach: Multi-Label Text Classification

Multi-Label Setup

This method preserved the independent nature of each label (Humorous, Sarcastic, Offensive, Vulgar) as binary values (0 or 1). Thus, each sample was associated with a four-element label vector, e.g., [1.0, 0.0, 1.0, 0.0].

Data Preparation

The dataset was again split into training (70%), validation (15%), and test (15%) sets, using stratified sampling based on the `Humorous` attribute to maintain class balance. The same PyTorch Dataset class setup was extended for multi-label scenarios.

Model Training

Fine-tuning was conducted under the following configuration:

- Loss Function: Binary Cross-Entropy with Logits Loss (`BCEWithLogitsLoss`).
- Optimizer: `AdamW` with a learning rate of 1.25×10^{-6}.
- Scheduler: Linear learning rate scheduler without warmup steps.
- Epochs: 15.
- Batch Size: 32.

- Hardware: GPU acceleration (if available).

During training:

- Forward pass was conducted with inputs and attention masks.
- Loss was computed independently for each label.
- Backpropagation and optimization were applied.
- Accuracy, precision, recall, and F1-scores were computed using the sample-based averaging method, appropriate for multi-label classification.

Confusion matrices were generated individually for each label to better visualize specific misclassifications.

4.2 Summary

This methodology aims to evaluate how well transformer models can understand the subtle and intertwined aspects of humor, sarcasm, offensiveness, and vulgarity in Bangla-English code-mixed text. The dual approach not only benchmarks conventional methods but also offers insights into task-specific model behavior in highly informal multilingual environments

Chapter 5

Result Analysis

5.1 Benchmarking BERT & LLMs

To evaluate the performance of transformer-based models on the MixSarc dataset, we benchmarked Banglish-BERT and Gemma-2B across four classification tasks: Humor, Sarcasm, Vulgar, and Offense. The evaluation metrics considered were Accuracy, Precision, Recall, and F1-Score.

Table 5.1: Performance Comparison of Banglish-BERT and Gemma-2B on MixSarc Dataset

Task	**Banglish-BERT**				**Gemma-2B**			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Humor	0.6232	0.623	0.8197	0.7080	0.6012	0.6007	0.8474	0.7031
Sarcasm	0.6569	0.3689	0.4222	0.3938	0.7287	0.4539	0.0944	0.1553
Vulgar	0.9509	0.5	0.1194	0.1928	0.9509	0.5	0.0299	0.0563
Offense	0.9508	0.125	0.0364	0.0563	0.9589	0	0	0

5.1.1 Banglish-BERT Performance

For Humor detection, Banglish-BERT achieved an accuracy of 0.6232 and an F1-score of 0.708, with recall reaching 0.8197. This indicates that the model is highly sensitive in identifying humorous instances, although precision remains moderate.

In the case of Sarcasm detection, the performance was comparatively weaker, with an accuracy of 0.6569 and an F1-score of 0.3938. The relatively low precision (0.3689)

and recall (0.4222) highlight the inherent difficulty of sarcasm detection in Bangla-English mixed text.

For Vulgar and Offense detection, Banglish-BERT achieved very high accuracies (0.9509 and 0.9508 respectively). However, this performance is partly due to the class imbalance in the dataset, as reflected in the low F1-scores (0.1928 and 0.0563) and recall values (0.1194 and 0.0364). This suggests that while the model can classify the dominant class correctly, it struggles to detect minority-class instances effectively.

5.1.2 Gemma-2B Performance

For Humor detection, Gemma-2B yielded results similar to Banglish-BERT, with an accuracy of 0.6012 and a strong F1-score of 0.7031, supported by a high recall (0.8474). This consistency suggests that both models handle humor detection reasonably well.

In Sarcasm detection, Gemma-2B outperformed Banglish-BERT in terms of accuracy (0.7287 vs. 0.6569) and precision (0.4539 vs. 0.3689), but the F1-score (0.1553) and recall (0.0944) were much lower. This indicates that Gemma-2B is more conservative in predicting sarcasm, resulting in higher precision but at the cost of failing to recall most sarcastic cases.

For Vulgar and Offense detection, Gemma-2B showed very high accuracies (0.9509 and 0.9589), similar to Banglish-BERT. However, the F1-scores were even lower (0.0563 and 0), with near-zero recall for offense detection. This further confirms the effect of dataset imbalance, where models achieve inflated accuracy by predicting the dominant class while ignoring minority instances.

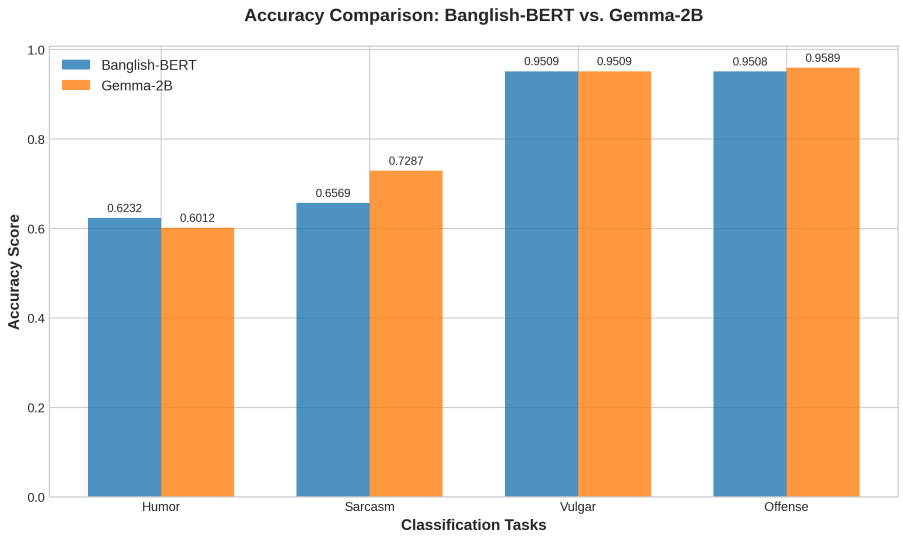

Figure 5.1: Accuracy comparison of Banglish-BERT and Gemma-2B across classification tasks.

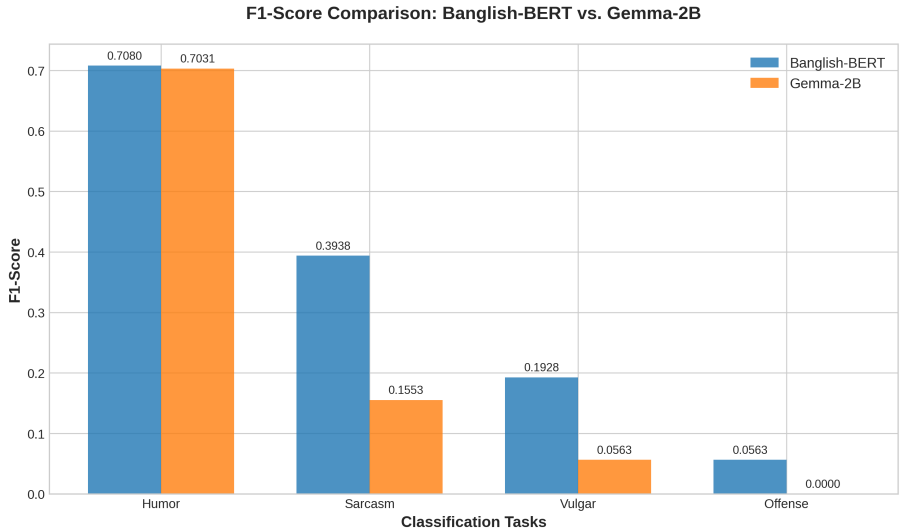

Figure 5.2: F1-score comparison of Banglish-BERT and Gemma-2B across classification tasks.

5.1.3 Comparative Insights

Both models perform well on Humor detection, with high recall values indicating good sensitivity to humorous content.

Sarcasm detection remains the most challenging task, as seen from the relatively low F1-scores across both models. Banglish-BERT achieves a more balanced trade-off between precision and recall, while Gemma-2B favors precision but misses most sarcastic cases.

In Vulgar and Offense detection, although accuracies appear high, the poor F1-scores reveal that minority-class instances are extremely rare or context-dependent. The near-zero F1-scores highlight that such nuanced samples are challenging to capture, underscoring the need for specialized datasets and targeted model training.

Overall, Banglish-BERT provides more balanced performance, particularly for sarcasm, while Gemma-2B shows stronger precision but fails to generalize well across minority classes.

This analysis underscores the importance of addressing rare or context-specific phenomena and designing specialized mechanisms for sarcasm detection in Bangla-English mixed datasets, as general-purpose transformer models struggle with nuanced and low-resource tasks.

5.2 Improving Sentiment Analysis

5.2.1 The Necessity of Sarcasm Detection: A BnSentMix Case Study

Beyond the benchmarking experiments, we conducted an additional evaluation by applying the MixSarc-finetuned BERT model on the BnSentMix corpus to validate the effectiveness of our approach in identifying sarcasm within Bangla-English mixed text.

5.2.2 Re-evaluating Negative Sentiments

The BnSentMix dataset contains a substantial portion of sentences labeled as negative sentiment. Traditionally, sentiment analysis models classify sarcastic sentences as negative due to their surface-level linguistic cues. However, sarcasm inherently carries an ironic tone that may not necessarily correspond to genuine negativity. Misclassifying these sentences distorts overall sentiment distributions and reduces the accuracy of downstream tasks such as opinion mining and customer feedback analysis.

Our experiment focused specifically on analyzing the negative sentiment class to determine how many of these sentences could in fact be sarcastic.

5.2.3 Findings

We observed that negatively classified sentences exhibited a 0.4213 probability of being sarcastic when processed through our fine-tuned BERT model. Out of a total of 6,172 sentences labeled as negative, this translates to approximately 2,600 sentences being potentially sarcastic rather than genuinely negative.

To visualize this key finding, Figure 5.3 shows the distribution of sarcastic versus genuinely negative sentences.

This result demonstrates that sarcasm constitutes a significant confounding factor in sentiment analysis for code-mixed Bangla-English text.

5.2.4 Implications

Improved Sentiment Reliability: By correcting sarcastic misclassifications, our approach enhances the granularity and reliability of sentiment analysis. This ensures

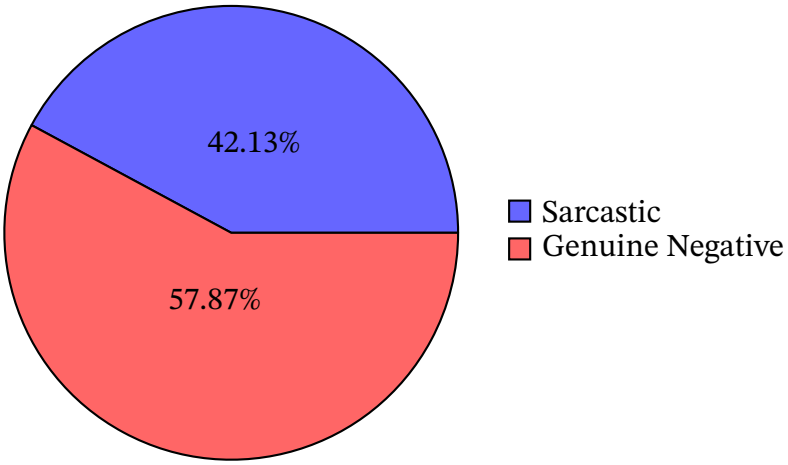

Figure 5.3: Proportion of Sarcastic vs. Genuine Negative sentences in the BnSentMix negative class.

that negative sentiment scores more accurately reflect genuinely negative opinions rather than ironic expressions.

5.2.5 Dataset-level Insight

The discovery that over 40% of negative-labeled content may actually be sarcasm highlights a major limitation in existing corpora. This calls for more nuanced annotation strategies that explicitly account for sarcasm in future dataset development.

5.2.6 Practical Applications

In domains such as social media monitoring, customer feedback analysis, and political discourse mining, the ability to differentiate sarcasm from true negativity is critical. For example, a sarcastic comment like "Na, na ekdomi poro na tumi. Ejonno 1st hou" should not be aggregated into negative sentiment statistics, as doing so may lead to incorrect business or policy decisions.

5.2.7 Conclusion of Insight

This experiment reinforces our novel contribution: sarcasm detection is not only an academic challenge but also a practical necessity for improving sentiment analysis in multilingual and code-mixed settings. By demonstrating that nearly half of the "negative" class in BnSentMix may actually be sarcasm, we provide concrete evidence of the impact and utility of MixSarc.

Chapter 6

Limitations and Scopes

While the proposed methodology establishes a strong foundation for humor, sarcasm, offensiveness, and vulgarity detection in Bangla-English code-mixed text, several avenues remain open for future research to further strengthen the system and address the limitations observed.

6.1 Dataset Expansion and Enhancement

- **Balanced Dataset:** Future efforts will focus on creating a more balanced dataset across all classes, particularly ensuring sufficient representation of humorous, sarcastic, offensive, and vulgar instances. This will help mitigate class imbalance issues during model training.
- **Increasing Offensive and Vulgar Samples:** Given that offensive and vulgar content tends to be underrepresented, dedicated scraping and annotation efforts will target sources rich in such language. Potential sources include:
  - Facebook public pages and groups
  - YouTube comment sections
  - Open-access forums and Bangladeshi meme pages
- **Advanced Filtering:** To improve the quality of extracted code-mixed data, filtering pipelines will be upgraded. Two potential techniques include:
  - Fine-tuned transformer-based language identification models.
  - Few-shot learning techniques using large language models (LLMs) like GPT-4 or Claude for quick labeling and content validation.

6.2 Comparative Analysis of LLMs and Traditional Transformers

Future work will involve a comparative study between:

- **Fine-tuned Transformer Models** (e.g., BERT, mBERT, XLM-R) and
- **Large Language Models (LLMs)** through few-shot or zero-shot prompting (e.g., GPT-3.5, GPT-4, Gemini).

The objective is to analyze performance variations when adjusting methodologies, such as:

- Training transformers from scratch versus using LLMs with dynamic prompts.
- Fine-tuning small transformers versus prompt engineering with LLMs.

This comparative analysis can provide insights into the trade-offs between traditional fine-tuning and emerging prompting paradigms in low-resource, code-mixed settings.

6.3 Summary

Overall, the future work aims to develop more comprehensive datasets, create robust, sarcasm- and offensiveness-aware models, explore LLM-based approaches, and introduce innovative evaluation metrics tailored for nuanced multilingual NLP tasks in code-mixed Bangla-English settings.[12]

Bibliography

[1] A. Abaskohi, A. Rasouli, T. Zeraati, and B. Bahrak, "Utnlp at semeval-2022 task 6: A comparative analysis of sarcasm detection using generative-based and mutation-based data augmentation," *arXiv preprint arXiv:2204.08198*, 2022.

[2] A. Aggarwal, A. Wadhawan, A. Chaudhary, and K. Maurya, ""' did you really mean what you said?": Sarcasm detection in hindi-english code-mixed data using bilingual word embeddings," *arXiv preprint arXiv:2010.00310*, 2020.

[3] S. Akrah, "Duluthnlp at semeval-2021 task 7: Fine-tuning roberta model for humor detection and offense rating," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 1196–1203.

[4] S. Alam, M. F. Ishmam, N. H. Alvee, M. S. Siddique, M. A. Hossain, and A. R. M. Kamal, "Bnsentmix: A diverse bengali-english code-mixed dataset for sentiment analysis," *arXiv preprint arXiv:2408.08964*, 2024.

[5] T. S. Apon, R. Anan, E. A. Modhu, A. Suter, I. J. Sneha, and M. G. R. Alam, "Banglasarc: A dataset for sarcasm detection," in *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, 2022, pp. 1–5.

[6] A. Chanda, D. Das, and C. Mazumdar, "Unraveling the english-bengali code-mixing phenomenon," in *Proceedings of the second workshop on computational approaches to code switching*, 2016, pp. 80–89.

[7] S. K. Lora et al., "Ben-sarc: A self-annotated corpus for sarcasm detection from bengali social media comments and its baseline evaluation," *Natural Language Processing*, vol. 31, no. 2, pp. 674–699, 2025.

[8] S. Oprea and W. Magdy, "Isarcasm: A dataset of intended sarcasm," *arXiv preprint arXiv:1911.03123*, 2019.

[9] M. N. Raihan, D. Goswami, and A. Mahmud, "Mixed-distil-bert: Code-mixed language modeling for bangla, english, and hindi," *arXiv preprint arXiv:2309.10272*, 2023.

[10] M. N. Raihan, D. Goswami, A. Mahmud, A. Anastasopoulos, and M. Zampieri, "Sentmix-3l: A bangla-english-hindi code-mixed dataset for sentiment analysis," *arXiv preprint arXiv:2310.18023*, 2023.

[11] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 704–714.

[12] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?" *LDV Forum*, vol. 22, no. 1, pp. 29–49, 2007.

[13] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, and M. Shrivastava, "A corpus of english-hindi code-mixed tweets for sarcasm detection," *arXiv preprint arXiv:1805.11869*, 2018.