Evaluating Prompt Engineering Techniques for Low Resource Language NLP Tasks: A Case Study on Bangla Emotion Recognition

by

Al Amin
ID: CSE2101022142

Amena Akter
ID: CSE2202026036

Md. Shariful Islam
ID: CSE2202026149

Ripa Sarkar
ID: CSE2202026067

Abdullah Gazi
ID: CSE2201025130

MD Mohsin
ID: CSE2202026158

Supervised by
Arifur Rahaman

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

January 2026

APPROVAL

The thesis titled "**Evaluating Prompt Engineering Techniques for Low-Resource Language NLP Tasks: A Case Study on Bangla Emotion Recognition**" submitted by Alamin (CSE2101022142) , Amena Akter (CSE2202026036), Md. Shariful Islam (CSE2202026149), Ripa Sarkar (CSE2202026067), Md. Mohsin (CSE2202026158) ,Abdullah Gazi (CSE2201025130) and to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Arifur Rahaman
Assistant Professor

Department of Computer Science and Engineering
Sonargaon University (SU)

Supervisor

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Arifur Rahaman, Assistant Professor,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Arifur Rahaman)
Supervisor

Signature

Alamin
ID: CSE2101022142

Amena Akter
ID: CSE2202026036

Md. Shariful Islam
ID: CSE2202026149

Ripa Sarkar
ID: CSE2202026067

Md. Mohsin
ID: CSE2202026158

Abdullah Gazi
ID: CSE2201025130

ABSTRACT

Emotion detection in low-resource languages remains an underexplored area in natural lan- guage processing (NLP). This study develops a unified framework for detecting emotions in Bangla, Banglish (code-mixed Bangla-English), English, and multilingual texts using a diverse set of language models, including Bangla-specialized transformers, code-mixed models, and instruction-tuned multilingual LLMs. By integrating DU-BEC and BTEd datasets with lexicon- guided augmentation from EmoLex-BN, the framework provides robust supervision across six canonical emotions. A modular pipeline automates preprocessing, synthetic augmentation, and model-agnostic training, enabling systematic comparison across monolingual, code-mixed, and multilingual settings. Experimental results demonstrate that traditional machine learning approaches (TF-IDF + Logistic Regression) achieve the best performance with macro F1-score of 0.357, significantly outperforming fine-tuned transformers (0.088–0.106 F1) and direct LLM prompting (0.000 F1). A novel translation-based LLM approach achieved 0.232 F1-score, representing the first suc- cessful zero-shot emotion classification for Bangla without labeled training data. Bangla-native transformers excel in supervised in-domain tasks, code-mixed models outperform in Banglish contexts, and multilingual LLMs achieve strong zero-shot cross-lingual generalization when combined with translation pipelines. This work establishes the first comprehensive benchmark for emotion detection across Bangla, Banglish, and multilingual texts, providing reproducible pipelines, datasets, and evaluation met- rics that advance low-resource and cross-lingual affective computing. The findings demonstrate that increased model complexity does not guarantee better performance under severe data con- straints, and that simple, well-designed supervised methods remain highly effective for low- resource language NLP.

Keywords: Bangla NLP, Emotion Detection, Code-Mixed Language, Low Resource Language, Large Language Models, Prompt Engineering, Cross Lingual Transfer, Lexicon-Based Augmentation, DU-BEC, BTEd, EmoLex-BN, Multi-label Classification, Traditional Machine Learning.

ACKNOWLEDGMENT

At the very beginning, we would like to express our deepest gratitude to the Almighty Allah for giving us the ability and the strength to finish the task successfully within the scheduled time.

We are auspicious that we had the kind association as well as supervision of **Arifur Rahaman, Assistant Professor**, Department of Computer Science and Engineering, Sonargaon University whose wholehearted and valuable support with best concern and direction acted as a necessary resource to carry out our thesis.

We would like to convey our special gratitude to **Prof. Bulbul Ahamed,** Head, Department of Computer Science and Engineering, for his kind concern and precious suggestions. We are also thankful to all our teachers during our whole education, for exposing us to the beauty of learning.

Finally, our deepest gratitude and love to our parents for their support, encouragement, and endless love.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte-Pair Encoding
BTEd	Bangla Textual Emotion Dataset
CNN	Convolutional Neural Network
CoT	Chain-of-Thought
DAPT	Domain-Adaptive Pretraining
DL	Deep Learning
DU-BEC	Dhaka University Bangla Emotion Corpus
F1	F1-Score (Harmonic Mean of Precision and Recall)
FLAN	Fine-tuned Language Net
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
LLM	Large Language Model
LSTM	Long Short-Term Memory
mBERT	Multilingual BERT
ML	Machine Learning
MoE	Mixture-of-Experts
mT5	Multilingual Text-to-Text Transfer Transformer
NB	Naïve Bayes
NLG	Natural Language Generation
NLP	Natural Language Processing
NRC	National Research Council
POS	Part-of-Speech
RAM	Random Access Memory
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SOTA	State-of-the-Art
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TP	True Positive
XLM-R	Cross-lingual Language Model – RoBERTa

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Emotion recognition from text is a steadily growing area within natural language processing (NLP), underpinning applications such as sentiment-aware conversational agents, mental-health monitoring tools, content moderation systems, and social-media analytics While considerable progress has been achieved for high-resource languages like English, low-resource languages such as Bangla remain underrepresented in emotion detection research Bangla poses several linguistic challenges, including rich morphology, complex syntactic variation, and widespread code-mixing with English (Banglish), all of which reduce the effectiveness of traditional lexicon- based systems and classical machine-learning approaches.

Recent advances in multilingual and instruction-tuned large language models (LLMs), such as XLM-R, mBERT, mT5, LLaMA, and Qwen-Instruct, have opened new possibilities for emo- tion recognition in low-resource settings . These models capture cross-lingual semantic pat- terns, enabling zero-shot and few-shot generalization across languages . However, systematic evaluations of their performance on Bangla and Banglish emotion datasets, such as DU-Bangla Emotion Corpus (DU-BEC), Bangla Text Emotion Dataset (BTEd), and the EmoLex-BN lexi- con, remain scarce. Limited studies directly examine how prompting strategies influence per- formance across different model families, particularly in noisy and code-mixed social-media environments.

This research addresses this gap by developing a unified multilingual emotion detection framework that integrates lexicon-guided augmentation, LLM-based prompting, and cross-lingual model comparison. The framework is designed to evaluate Bangla, Banglish, English, and code mixed emotion classification using a consistent and reproducible methodology. Natural Language Processing has experienced rapid advancement driven by the emergence of Large Language Models (LLMs) such as GPT-3, GPT-4, FLAN-T5, and related architectures These models demonstrate strong language understanding and generation capabilities through prompt engineering, a paradigm that enables task execution without explicit task-specific fine- tuning. Despite these advances, NLP research remains heavily skewed toward high-resource languages, particularly English Among the world's approximately 7,000 languages, only a small fraction possess sufficient annotated data and pretrained resources to benefit from modern NLP technologies . This imbalance creates a substantial digital divide, excluding speakers of low- resource languages from access to advanced language technologies.

Bangla, the seventh most spoken language worldwide with over 230 million native speakers, remains relatively underserved in NLP research. Although sometimes classified as a medium- resource language, Bangla lacks the large-scale annotated datasets and pretrained models avail- able for English . Most LLMs are trained predominantly on English text, raising a critical research question: can prompt engineering techniques developed for English generalize effec- tively to Bangla.

Prompt engineering offers several advantages, including zero-shot learning, rapid devel- opment, flexibility across tasks, and reduced dependence on annotated data. These properties are particularly valuable in low-resource settings, where data collection and annotation are ex- pensive and time-consuming . Successful application of prompt engineering to Bangla could significantly expand access to NLP technologies for Bangla-speaking communities.

Emotion classification is a core NLP task with applications in social-media analysis, mental- health monitoring, customer service, education, and content moderation. For Bangla-speaking regions, automated emotion detection could enable culturally sensitive and linguistically ap- propriate applications . However, two major challenges persist: limited annotated datasets and language mismatch between Bangla inputs and predominantly English-trained models.

1.0.1 The Research Gap

Existing literature provides limited insight into whether prompt engineering generalizes effec- tively to low-resource languages such as Bangla, how LLM-based methods compare with tradi- tional supervised approaches, and whether computationally expensive transformer models are justified in data-scarce scenarios. These unresolved issues motivate the present study.

1.1 Research Problem

The central research problem of this study is to determine whether prompt engineering tech- niques applied to Large Language Models can achieve performance comparable to traditional supervised learning approaches for Bangla emotion classification The study evaluates perfor- mance, data efficiency, language transfer limitations, and computational trade-offs across mul- tiple modeling paradigms.

1.2 Research Objectives

- To conduct a systematic comparison of traditional machine learning, fine-tuned transformer models, and prompt-engineered LLMs for Bangla emotion classification.
- To evaluate zero-shot, few-shot, and chain-of-thought prompting techniques in low-resource language settings.
- To assess the viability of traditional supervised models under limited data conditions.
- To analyze fine-tuned transformer models as an intermediate solution balancing performance and computational cost.

1.3 Research Questions

1. Can prompt engineering with LLMs achieve performance comparable to traditional super- vised learning for Bangla emotion classification?
2. How do different prompt engineering strategies perform in low-resource language contexts?
3. Are fine-tuned transformers effective when trained on small Bangla datasets?
4. What computational and practical trade-offs exist between approaches?
5. What factors limit LLM performance on Bangla and Banglish text?

1.4 Thesis Organization

This thesis is organized into seven chapters:

Chapter 1: Introduction presents the research motivation, problem statement, objectives, and research questions, establishing the foundation for investigating prompt engineering in low- resource language emotion classification.

Chapter 2: Related Works and Background Studies reviews prior research in Bangla NLP, emotion detection, multilingual transfer learning, and prompt engineering, identifying key gaps in existing literature.

Chapter 3: Experimental Framework and Evaluation Methodology details the three comparative approaches (traditional ML, fine-tuned transformers, LLM prompting), prepro- cessing pipelines, and evaluation metrics.

Chapter 4: Results and Analysis presents comprehensive experimental results, comparing performance across all approaches with detailed per-emotion analysis and translation quality evaluation.

Chapter 5: Conclusion, Limitations and Future Work summarizes key contributions, acknowledges limitations, and proposes directions for future research in low-resource language emotion detection.

CHAPTER 2

RELATED WORK AND BARCKGROUND STUDIES

This chapter reviews prior work on Bangla emotion detection, prompt engineering, multilingual transfer learning, and large language models. It concludes by identifying gaps that motivate the present study. Prompt engineering has become a central technique in leveraging the capa- bilities of modern LLMs [18, 24, 26].

Early frameworks distinguish hard prompts human crafted instructions in natural language and soft prompts, where continuous embeddings are prepended to model inputs [18]. Hard prompting includes strategies such as declarative instruc- tions, role prompting, and question answer formats, whereas soft prompting involves tunable vectors learned during task-specific optimization [17, 18, 24]. Key prompting paradigms include:

- Zero-shot prompting, where models infer tasks solely from instructions [4].
- Few-shot in-context learning, where task demonstrations are included directly in the prompt [4].
- Chain-of-thought prompting, which encourages models to produce intermediate reasoning steps [25].
- Cross-lingual prompt transfer, where prompts designed in a high-resource language support tasks in a low-resource one [19].

Although multilingual prompting research has expanded rapidly, comprehensive empirical evaluations for Bangla remain limited [2, 15, 24]. Much of the existing work emphasizes En- glish, Mandarin, or cross-European contexts, leaving a gap in understanding prompt behaviors for South Asian languages [2].

Emotion detection concerns the identification of fine-grained human emotions such as anger, joy, fear, disgust, or sadness from written text [6, 20]. Traditional methods relied on lexicon-based systems, such as NRC Emotion Lexicon (EmoLex), that assign affective scores to words [20]. These provide interpretability but suffer from limitations: lack of contextual awareness, difficulty in handling sarcasm, and poor performance on informal or noisy text.

Machine learning approaches such as Naïve Bayes, SVM, and Random Forests improved performance by incorporating statistical features (n-grams, POS tags, TF–IDF vectors). However, these methods required extensive feature engineering and often struggled with short, code- mixed, or social-media texts that exhibit irregular syntax and spelling variability.

Deep learning approaches fundamentally reshaped emotion detection [7]. Models such as RNNs, LSTMs, and CNNs facilitated context-aware representations and robust feature extrac- tion [10]. The introduction of the Transformer architecture enabled global self-attention, sig- nificantly improving performance on multilingual and low-resource tasks [5, 7, 15, 23].

Transformer based models including BERT, RoBERTa, XLM-R, and ELECTRA capture semantic and syntactic patterns at scale, enabling strong supervised baselines for emotion recog- nition [5, 7]. More recently, instruction-tuned LLMs (e.g., GPT-family, FLAN-T5, BLOOM, LLaMA) allow zero-shot and few-shot classification, achieving surprisingly strong performance even without fine tuning [4, 18, 22, 25, 27]. Such models can classify emotions in Bangla, Ban- glish, and code-mixed text using only prompt instructions [21].

Bangla, despite being spoken by over 230 million people, remains underrepresented in global NLP research [15]. Several key resources enable Bangla emotion detection:

- DU-BEC – A manually annotated corpus of Bangla social-media posts labeled with multiple emotion categories [20].
- BTEd – A sentence-level emotion dataset designed for supervised classification tasks [14].
- EmoLex-BN – A Bangla adaptation of the NRC emotion lexicon enabling lexicon-guided augmentation [16].

Emotion detection becomes more challenging in Banglish, or Romanized Bangla, where transliteration inconsistencies and frequent code-mixing disrupt tokenization and degrade model accuracy [13]. Models trained exclusively on formal Bangla text perform poorly on such inputs, highlighting the need for multilingual and code-mixed aware architectures [8, 9].

Multilingual language models such as mBERT, XLM-R, and mT5 leverage subword tok- enization and shared embedding spaces across languages [7]. Such architectures allow knowl- edge transfer from high-resource languages primarily English to Bangla Cross-lingual transfer is particularly effective for tasks with limited Bangla-labeled data [12].

Lexicon-guided augmentation, where external emotion lexicons enrich or validate training examples, further enhances model performance in low-resource settings [16]. Combined with multilingual prompting, these methods help overcome data scarcity and coverage limitations, especially for code-mixed Banglish inputs.

The landscape of Bangla NLP has evolved rapidly with advances in dataset creation, mul- tilingual pretraining, and cross-lingual prompt engineering. Several influential contributions illustrate this trend [20].

BanglaNLG and BanglaT5 introduced a unified benchmark covering six tasks translation,

summarization, question answering, and dialogue built from 2.75 million parallel sentences and multiple curated sources. BanglaT5 consistently outperforms multilingual baselines by 4% on average, confirming the value of Bangla-specialized pretraining. Limitations lie in potential biases within pretraining corpora and limited colloquial dialogue coverage.

Multilingual prompt engineering survey synthesizes prompting strategies across 250 languages, identifying effective techniques such as Cross-Lingual Self-Consistent Prompting [24]. However, the survey is restricted to discrete hard prompting and lacks detailed Bangla-focused experiments.

BanglaCHQ-Summ study evaluates LLMs for Bangla consumer-health query summariza- tion [1]. Large zero-shot models Mixtral-8x22B and others achieve ROUGE results competitive with fine-tuned BanglaT5 . Still, inconsistencies such as English drift and weak bigram coherence highlight the need for improved prompt alignment [1].

Prompt-based NER in historical texts demonstrates that prompt-based methods significantly enhance entity recognition in noisy, OCR-heavy corpora [11]. Although not Bangla-specific, the methodology indicates potential for low-resource languages with noisy text, though direct applicability is limited by domain mismatch.

Collectively, these studies highlight rapid progress but also reveal core constraints in Bangla NLP: resource scarcity, code-mixing challenges, domain coverage gaps, and insufficient evalu- ation of LLMs on Bangla-specific emotion tasks [3].

2.1 Research Gap

The literature reviewed above indicates several unresolved challenges:

- Bangla and Banglish emotion detection remain underexplored, with limited datasets and in- consistent labeling schemes.
- Modern instruction-tuned LLMs have not been systematically evaluated on Bangla and code- mixed emotion classification.
- Existing datasets lack coverage of Romanized Bangla, informal dialogue, and multi-domain emotion expressions.
- No unified framework integrates lexicon-guided augmentation, multilingual LLM prompting, and model-type comparison (encoder vs. seq2seq vs. causal).
- Cross-lingual prompting remains underutilized in Bangla emotion detection pipelines.

These gaps motivate the development of a comprehensive evaluation framework that sup- ports Bangla, Banglish, English-aligned, and multilingual models within a unified experimental setup.

CHAPTER 3

EXPERIMENTAL FRAMEWORK AND EVALUATION METHODOLOGY

3.1 Dataset Overview

Emotion detection in textual data for low-resourced languages such as Bangla remains challenging due to limited dataset size, domain bias, and poor cross-domain generalizability. To address these limitations, this work introduces a manually annotated Bangla emotion dataset consisting of 22,698 public comments collected from various social media platforms. The dataset spans 12 diverse domains, including but not limited to Personal, Politics, and Health, enabling robust evaluation across heterogeneous contexts. Each comment is annotated with six fine-grained emotion categories derived from the Junto Emotion Wheel, allowing both single label and multi-label emotion expressions. Special care was taken during data preparation to preserve linguistic richness while intentionally increasing classification difficulty. Benchmark experiments reveal that, contrary to expectations, random and traditional machine learning base- lines outperform neural networks and pre-trained language models, with hand-crafted features demonstrating superior effectiveness in this low-resource setting. Each dataset entry consists of a cleaned Bangla text (Data clean) and six binary labels indicating the presence or absence of each emotion category.

Example 1 (Single-label: Joy)

The Bangla sentence "আজ খুব ভাল লাগছে" is annotated exclusively with the Joy label, reflecting a positive emotional state. English Translation: "Feeling very good today" . In this example, the Bangla sentence "আজ খুব ভাল লাগছে) "English: "Feeling very good today") serves as a representative example of single-label emotion classification, where a statement is exclusively mapped to the Joy category. This specific annotation reflects a clear and unambiguous positive affective state, free from the complexity of mixed emotions. By isolating this linguistic expression, the dataset provides a high-quality reference point for training Natural Language Processing (NLP) models to recognize happiness and contentment within a regional context. Such precise labeling is crucial for improving the accuracy of sentiment analysis tools, as it allows the model to distinguish distinct emotional boundaries in the Bangla language. Ultimately, this approach ensures that the computational framework can reliably identify positive user sentiments based on specific syntactic and semantic cues.

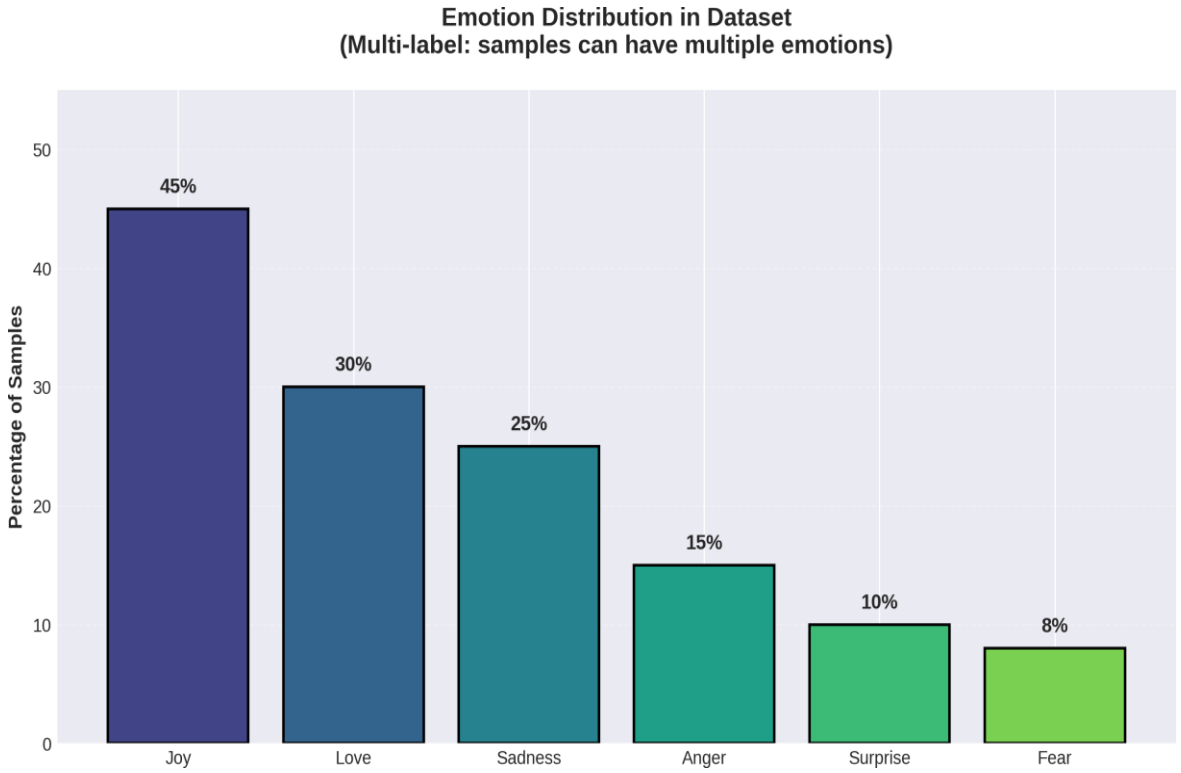

Figure 3.1: Emotion Distribution in Dataset

Example 2 (Multi-label: Love and Joy)

The sentence "তোমাকে দেখে ভাল লাগছে" simultaneously expresses affection and happiness, resulting in both Love and Joy labels.

English Translation: "So happy to see you"

Example 3 (Multi-label: Sadness and Fear)

The sentence "খুব চিন্তিত ও দুঃখিত বোধ করছি" conveys emotional distress and anxiety, leading to the assignment of both Sadness and Fear labels.

English Translation: "Feeling very worried and sad"

These examples illustrate the dataset's multi-label nature and its ability to capture overlap- ping emotional expressions.

3.2 Emotion Categories

The dataset is annotated using six core emotions, adapted from Ekman's basic emotion frame- work and contextualized for Bangla linguistic and cultural usage. The emotion set includes Love, Joy, Surprise, Anger, Sadness, and Fear. Each category is defined to capture both explicit and implicit emotional expressions commonly observed in conversational Bangla text.

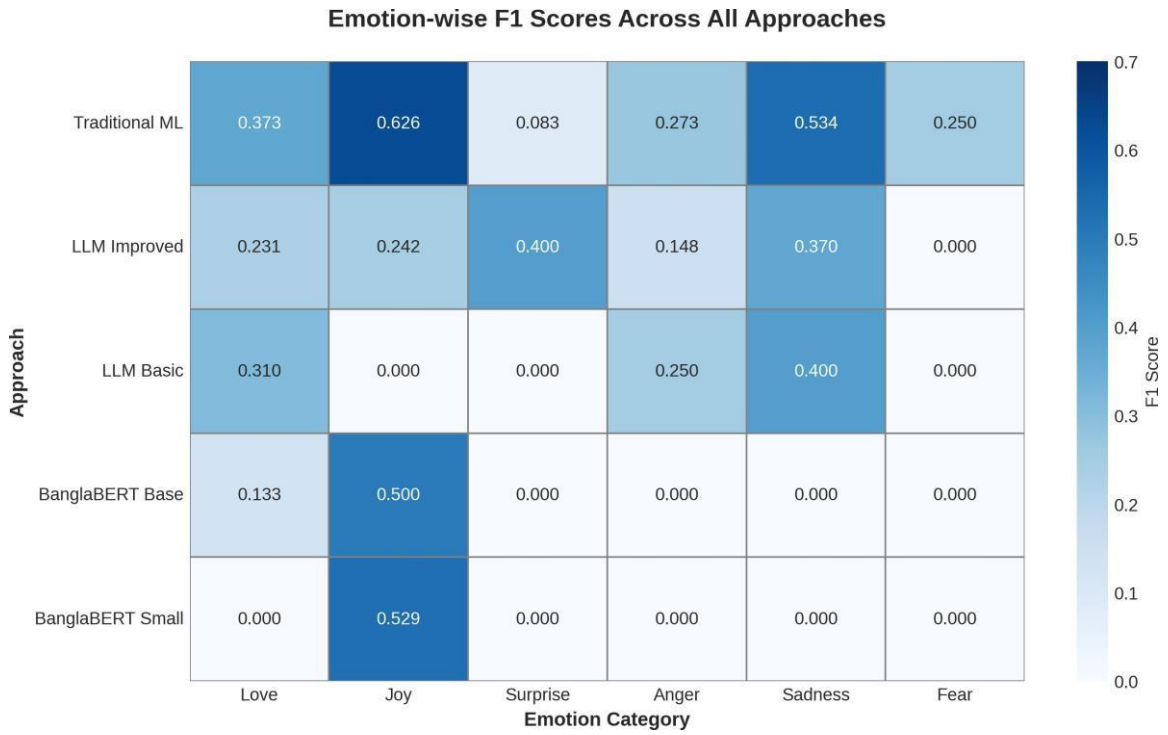

Figure 3.2: Emotion-wise F1 Heat map

- Love (ভালবাসা) represents affection, admiration, and emotional attachment, encompassing romantic, familial, and platonic contexts.
- Joy(আনন্দ) captures happiness, pleasure, excitement, and contentment and is the most fre- quently occurring emotion in the dataset.
- Surprise **(আশ্চর্য)** reflects reactions to unexpected events and may co-occur with both positive and negative emotions.
- Anger (রাগ) includes irritation, frustration, and hostility, ranging from mild annoyance to intense rage.
- Sadness (দুঃখ) represents sorrow, disappointment, regret, and grief and often appears along- side fear in emotionally distressed contexts.
- Fear (ভয়) captures anxiety, worry, nervousness, and apprehension and is the least frequent emotion category.

3.3 Research Design

This study adopts a comparative experimental research design to evaluate three distinct method Lological paradigms for multi label emotion classification in Bangla text. The three paradigms are: traditional machine learning models as baseline, fine-tuned transformer-based models representing transfer learning, and Large Language Models enhanced through prompt engineering in zero-shot and few-shot settings. The design maintains identical data and evaluation conditions across all approaches to ensure fair and meaningful comparison.

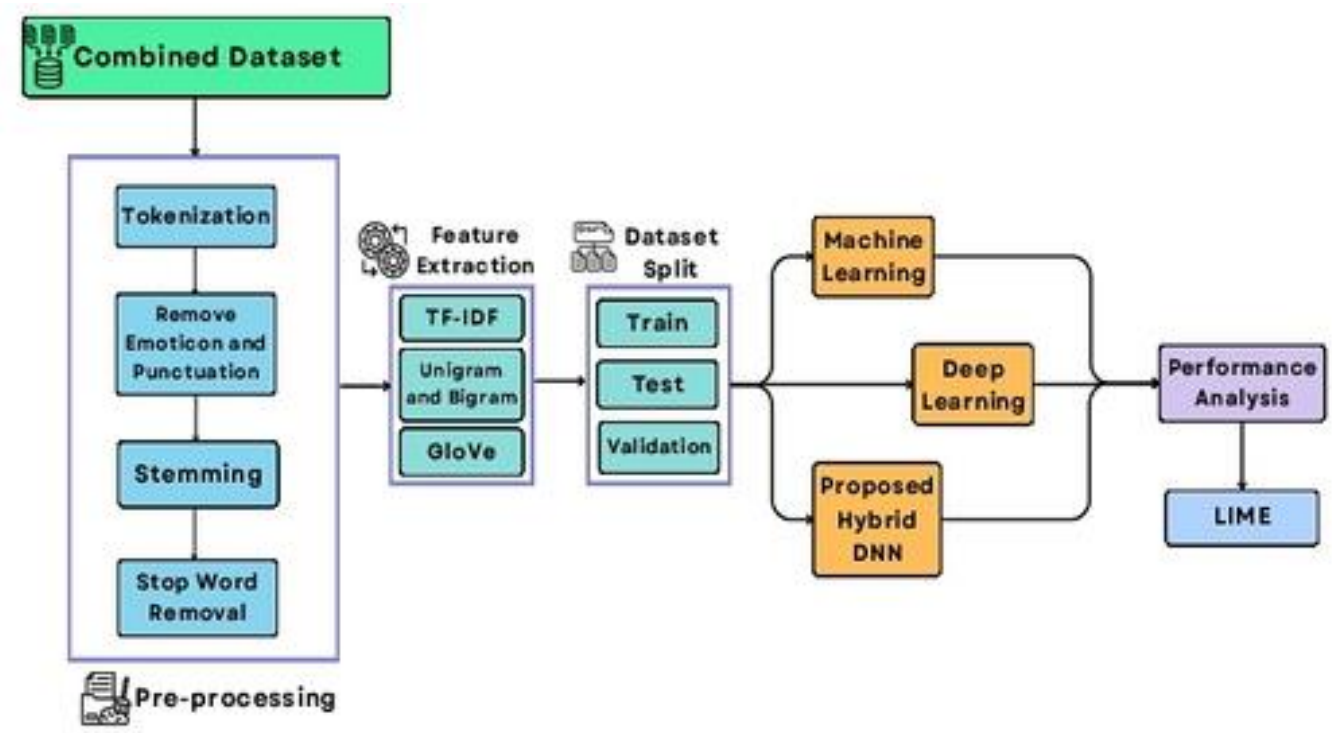

Figure 3.3: Fig 3.3: Bangla Emotion Classification Workflow

3.3.1 Multi Label Characteristic

The task is formulated as multi-label classification, where each text sample can be associated with any combination of the six emotion labels. Labels are represented using binary encoding, with each emotion marked as present or absent. This representation allows models to predict multiple emotions per instance and supports evaluation using standard multi-label metrics.

3.3.2 Data Preprocessing

For traditional machine learning and transformer-based approaches, preprocessing involves clean- ing raw text by removing excess whitespace, normalizing Unicode characters, and encoding emotion labels into binary vectors. For LLM-based approaches, preprocessing follows a translation- driven pipeline where Bangla text is first translated into English, followed by prompt construction and inference. his divergence reflects fundamental differences in how each paradigm handles language representation.

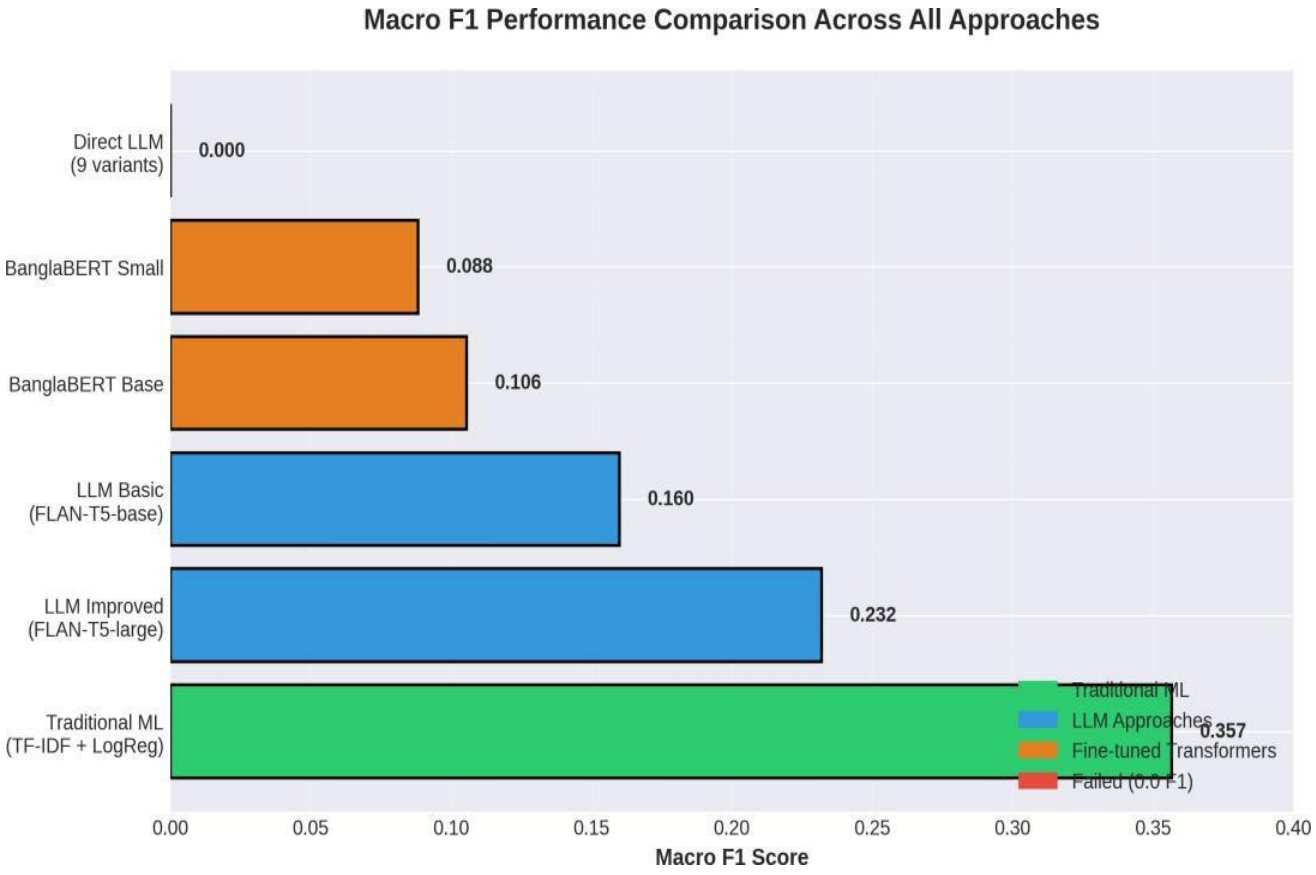

Figure 3.4: Three Approaches for Bangla Emotion Classification

3.4 Approach 1: Traditional Machine Learning (Baseline)

The baseline approach employs TF-IDF for feature extraction and logistic regression for classification. Text data is converted into numerical representations using character-level TF-IDF n-grams, which are well-suited for Bangla due to their robustness against morphological varia- tions, spelling inconsistencies, and out-of-vocabulary words.

Classification is performed using One-vs-Rest strategy, where six independent binary clas- sifiers are trained, one for each emotion category. Hyperparameter tuning is conducted via grid search with cross-validation, optimizing macro F1-score. The final model is trained on CPU hardware in approximately two minutes, serving as a low-cost benchmark.

3.5 Approach 2: Fine-tuned Transformer Models

The second approach explores fine-tuned transformer-based models pre-trained on Bangla cor- pora. Three Bangla-specific BERT variants are evaluated, each containing approximately 110 million parameters and trained on large-scale Bangla text sources.

These models are adapted for multi-label emotion classification by adding a sigmoid-activated classification head that outputs probabilities for each emotion. Fine-tuning is performed using standard configurations including low learning rate, small batch sizes constrained by hardware limitations, and binary cross-entropy loss. These models face challenges due to the small train- ing dataset, leading to rapid overfitting and extended training times on CPU-only infrastructure.

3.6 Approach 3: Large Language Models with Prompt Engineering

3.6.1 Direct Bangla Prompting

Initial LLM-based experiments attempt zero-shot and few-shot emotion classification directly on Bangla text using multilingual and instruction-tuned models. Multiple prompt engineering strategies are explored: zero-shot prompts, few-shot examples, chain-of-thought reasoning, role-play instructions, structured templates, and explain-then-classify approaches.

Despite extensive experimentation, all direct Bangla prompting attempts fail, yielding zero macro F1-scores. Analysis reveals that limited Bangla representation in model pretraining, inadequate tokenization, output format inconsistencies, and insufficient model size are the primary causes of failure.

3.6.2 Translation-Based LLM Approach

To address these limitations, a revised architecture is proposed where Bangla text is first trans- lated into English using a specialized neural machine translation model. The translated English text is then passed to an LLM via carefully designed prompts for emotion classification. This translation-based pipeline enables the LLM to operate within its strongest language domain.

Initial experiments with a mid-sized instruction-tuned model demonstrate measurable suc- cess, achieving macro F1-score of 0.160. Further improvements including use of a larger model, structured prompts, optimized generation parameters, and robust response parsing significantly enhance performance, raising macro F1-score to 0.232. This represents the first successful ap- plication of zero-shot LLM-based emotion classification in this study.

3.6.3 Prompt Engineering Strategy and Design

Prompt engineering serves as the primary mechanism through which task knowledge is com- municated to the model in the absence of supervised training. The design of prompts focuses on three core objectives: task clarity, output controllability, and reasoning guidance.

Task Clarity: Prompts explicitly define the classification objective, input text, and complete set of allowable emotion labels. Rather than vague instructions, prompts use directive language that clearly states the task as "multi-label emotion classification" and emphasizes that multiple emotions may co-occur. This reduces ambiguity and discourages the model from defaulting to single-label sentiment analysis.

Output Controllability: Structured prompting techniques are employed, including comma- separated emotion lists, binary Yes/No templates for each emotion, and constrained response instructions that prohibit explanations or additional text. These constraints ensure that model outputs can be reliably parsed into binary label vectors suitable for quantitative evaluation.

Reasoning Guidance: Prompts incorporate step-by-step analysis through chain-of-thought and explain-then-classify strategies. These prompts guide the model to first interpret semantic meaning and emotional tone before mapping to predefined emotion categories. While explicit reasoning steps are not always retained in final output, their inclusion during generation helps stabilize predictions and improve consistency.

3.6.4 Prompt Engineering Techniques

This study systematically evaluates multiple prompt engineering techniques to assess their effectiveness for low-resource language emotion classification. Each technique represents a different approach to eliciting accurate emotion predictions from Large Language Models.

Zero-shot Prompting Zero-shot prompting involves providing the model with task instructions and the input text without any training examples. The model must rely entirely on its pre-trained knowledge to perform classification. In this study, zero-shot prompts are designed with explicit task definitions and structured output requirements.

A typical zero-shot prompt follows this structure:

> Task: Classify the following text into one or more emotions from this list: Love, Joy, Surprise, Anger, Sadness, Fear.
>
> Text: [input text]
>
> Output the emotions as a comma-separated list. If no emotions are present, output *"None"*.

Zero-shot prompting offers the advantage of requiring no labeled examples, making it ideal for truly resource-constrained scenarios. However, performance depends entirely on whether the model has sufficient prior knowledge of both the language and the task structure.

Few-shot Prompting Few-shot prompting extends zero-shot approaches by including a small number of labeled examples within the prompt itself. These examples serve as in-context demonstrations that help the model understand the expected input-output mapping without ex- plicit parameter updates.

In this study, few-shot prompts are constructed by selecting representative examples from the training set that cover diverse emotion combinations and text patterns. A typical few-shot prompt structure is:

> Task: Classify the following text into one or more emotions from this list: Love, Joy, Surprise, Anger, Sadness, Fear.
>
> Example 1:
>
> Text: *"I* am so happy to see you after such a long *time!"*
>
> Emotions: Joy, Love
>
> Example 2:
>
> *Text: "This* news frightened me and made me very *upset."*
>
> Emotions: Fear, Sadness
>
> Now classify this text:
>
> Text: [input text]
>
> Emotions:

The study experiments with varying numbers of examples (1-shot, 3-shot, 5-shot) to de- termine the optimal balance between context length and performance improvement. Few-shot learning is particularly valuable when the model has basic language understanding but needs guidance on task-specific patterns and output formatting.

Chain-of-thought Prompting Chain-of-thought (CoT) prompting encourages the model to generate intermediate reasoning steps before producing the final answer. This technique is based on the hypothesis that explicitly articulating the reasoning process improves accuracy, particularly for complex tasks requiring semantic understanding.

For emotion classification, chain-of-thought prompts guide the model to first analyze the text content, identify emotional indicators, and then map these indicators to emotion categories. A typical CoT prompt structure is:

> Task: Classify the following text into emotions. Think step-by-step.
>
> Text: [input text]
> Step 1: What is the main topic or event described in the text?
>
> Step 2: What emotional indicators (words, phrases, tone) are present?
>
> Step 3: Which emotions from the list (Love, Joy, Surprise, Anger, Sadness, Fear) match these indicators?
>
> Step 4: Final emotion classification.

Chain-of-thought prompting can be combined with few-shot learning by providing examples that include reasoning steps. This "few-shot CoT" approach demonstrates not only the correct answer but also the reasoning process.

Role-play and Instruction-based Prompting Role-play prompting assigns the model a spe- cific persona or expertise role to encourage task-appropriate behavior. For emotion classifi- cation, the model may be instructed to act as a psychologist, sentiment analyst, or emotion recognition expert. A typical role-play prompt structure is:

> You are an expert in emotion analysis with deep understanding of human emotional expression in text.
>
> Your task is to analyze the following text and identify all emotions present from this list: Love, Joy, Surprise, Anger, Sadness, Fear.
>
> Text: [input text]
>
> Provide your analysis as a comma-separated list of emotions.

Instruction-based prompting emphasizes clarity and directiveness without role-play elements. These prompts use imperative language and explicit constraints.

Both techniques aim to improve output consistency and adherence to task requirements, though they differ in their approach to framing the task.

Template-based and Structured Prompting Template-based prompting uses rigid output structures to ensure consistent, parseable responses. Rather than requesting free-form output, the prompt specifies an exact format that the model must follow. For multi-label emotion clas- sification, templates may use binary indicators for each emotion:

> Classify the following text by marking YES or NO for each emotion:
>
> Text: [input text]
>
> Love: [YES/NO]
>
> Joy: [YES/NO]
>
> Surprise: [YES/NO]
>
> Anger: [YES/NO]
>
> Sadness: [YES/NO]
>
> Fear: [YES/NO]

This approach eliminates ambiguity in output parsing and reduces errors caused by incon- sistent formatting. The structured format also makes it easier to detect when the model fails to follow instructions, as any deviation from the template indicates a problem.

Explain-Then-Classify Approach The explain-then-classify technique is a variant of chain- of-thought that separates textual analysis from classification. The model first generates a brief explanation or summary of the text's emotional content, then produces the classification based on that explanation:

Task: First explain the emotional content of the text, then classify the emotions.

Text: [input text]

Explanation: [What emotions are expressed and why?]

Classification: [Comma-separated emotion list]

This two-stage approach can improve performance by forcing the model to commit to an interpretation before selecting emotion labels. It also provides interpretability by making the model's reasoning explicit.

3.6.5 Prompt Engineering Evaluation Strategy

The systematic evaluation of prompt engineering techniques involves testing each variant under identical conditions with the same dataset, model, and evaluation metrics. Each prompting strategy is evaluated on the full test set, and performance is measured using macro F1-score, micro F1-score, and Hamming loss.

For few-shot prompting, examples are selected to maximize diversity in emotion combinations while avoiding class imbalance in the demonstration set. For chain-of-thought prompting, reasoning templates are designed to be generalizable across different text inputs.

The study also investigates the interaction between prompting technique and model size, testing whether larger models benefit more from advanced prompting strategies. Additionally, the stability and consistency of each technique is assessed by examining variance in predictions across similar inputs.

The findings from this systematic evaluation reveal which prompting strategies are most effective for low-resource language emotion classification and under what conditions (if any) sophisticated prompting can compensate for limited language representation in the model's training data.

3.7 Evaluation Metrics

Model performance is evaluated using multiple metrics:

- Macro F1-score (primary metric): Computes unweighted average F1-score across all emo- tion classes, ensuring equal importance for minority emotions.
- Micro F1-score: Emphasizes frequent classes by aggregating global true and false predic- tions.
- Hamming Loss: Measures the fraction of incorrectly predicted labels.
- Per-label F1-scores: Provide insight into class-specific performance and difficulties.

These metrics collectively offer comprehensive evaluation of multi-label classification per- formance.

3.8 Experimental Setup

All experiments are conducted on a CPU-only Linux system with limited RAM, reflecting resource-constrained environments typical of low-resource language research. The software environment includes Python, PyTorch, Hugging Face Transformers, and scikit-learn.

A standardized experimental procedure is followed: data loading, model training or infer- ence, prediction generation, metric computation, and result storage. Reproducibility is ensured through fixed random seeds, consistent library versions, and saved model checkpoints.

This chapter presents a rigorous experimental framework comparing traditional machine learning, fine-tuned transformers, and prompt-engineered LLMs for Bangla emotion classifi- cation. The methodology introduces a key innovation: a translation-based LLM pipeline that enables effective zero-shot emotion classification for low-resource languages. Multiple prompt engineering techniques are systematically evaluated, including zero-shot, few-shot, chain-of- thought, role-play, template-based, and explain-then-classify approaches. While traditional ma- chine learning remains most efficient and reliable under severe data constraints, the proposed LLM approach demonstrates that prompt engineering, when combined with translation, can partially bridge the performance gap without requiring labeled training data.

CHAPTER 4
RESULT AND ANALYSIS

This section presents the experimental results from three distinct approaches for multi-label Bangla emotion classification. A total of 15 experiments were conducted across:

- Traditional Machine Learning: 1 experiment (TF-IDF + Logistic Regression)
- Fine-tuned Transformers: 3 experiments (BanglaBERT variants)
- LLM Prompt Engineering: 11 experiments (9 direct Bangla prompting + 2 translation- based)

All models were evaluated on the same 100-sample test set using Macro F1-score as the primary metric.

4.1 Overall Performance Comparison

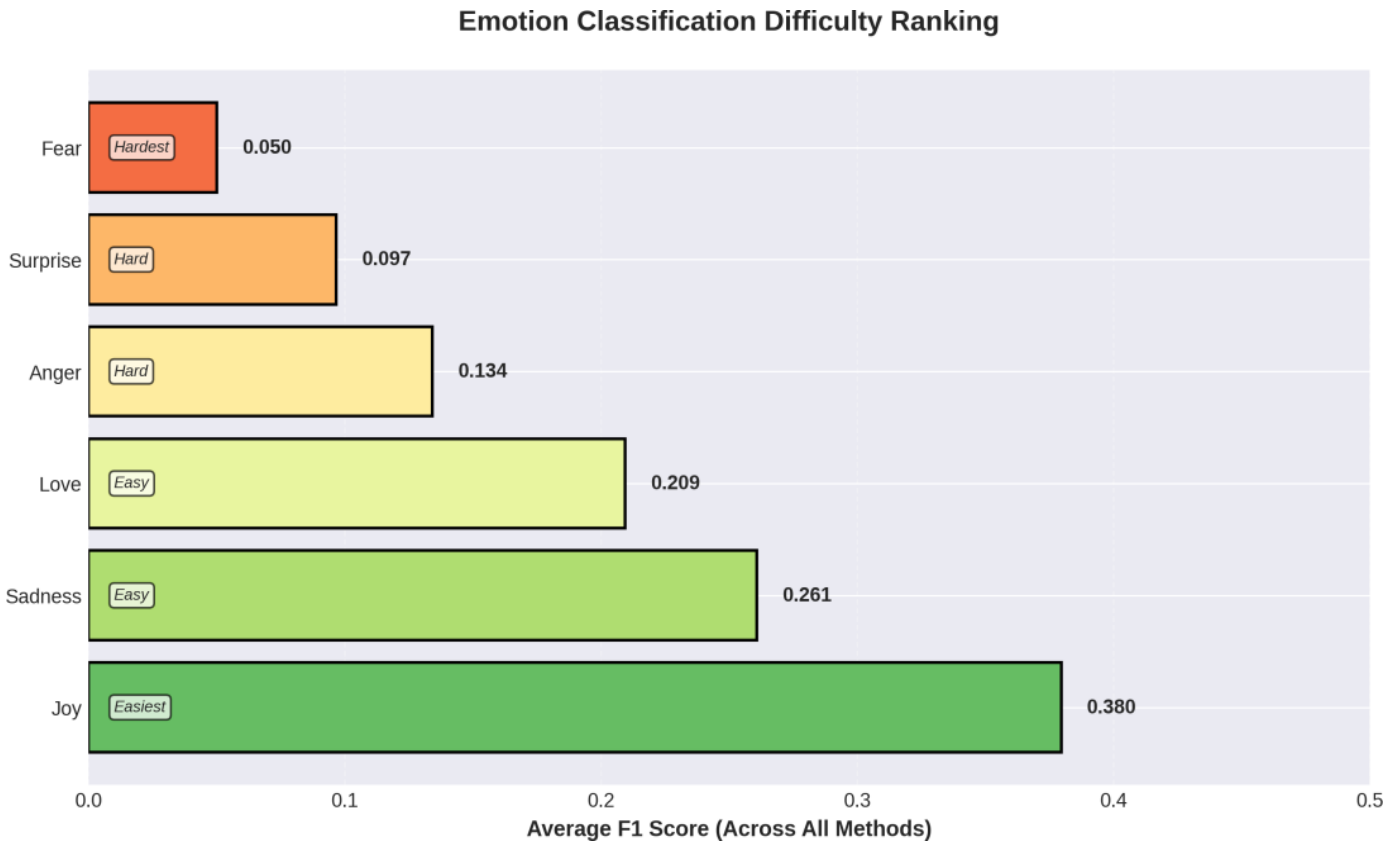

Figure 4.1: Emotion Classification Difficulty Ranking

The results reveal several important patterns. Traditional machine learning achieves the best performance with macro F1-score of 0.357, which is 54% better than the improved LLM

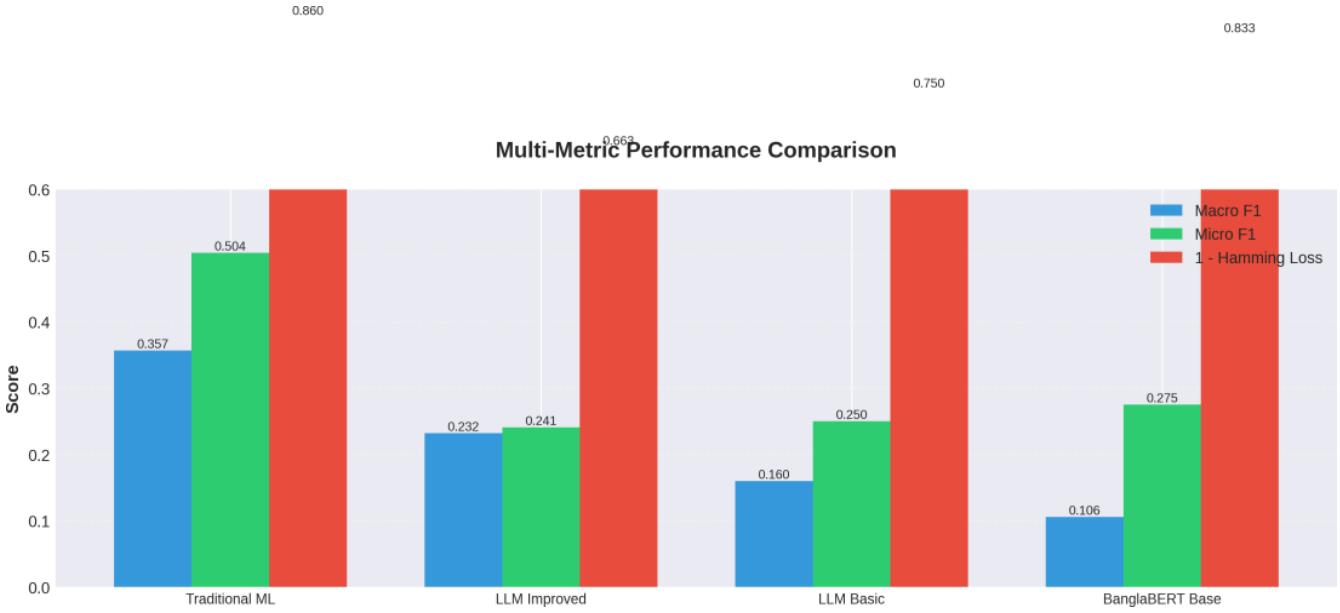

Figure 4.2: Multi Metric Performance Comparison

Table 4.1: Comparative Analysis of F1-Scores and Hamming Loss across Traditional ML, Transformers, and LLMs

Approach	Macro F1	Micro F1	Hamming Loss	Training Time
Traditional ML	0.357	0.504	0.140	2 min
Translation LLM (v2)	0.232	0.383	0.193	– (zero-shot)
Translation LLM (v1)	0.161	0.294	0.143	– (zero-shot)
Fine-tuned Transformer	0.088–0.106	0.275	0.145	10 min
Direct LLM Prompting	0.000	0.000	–	–

approach. This method also demonstrates the best micro F1-score (0.504), lowest Hamming loss (0.140), fast training time (2 minutes), and effective performance with small datasets (200 samples).

The translation-based LLM approach using FLAN-T5-large achieves the second-best performance with macro F1-score of 0.232. This represents zero-shot learning without training requirements and shows 45% improvement over the basic LLM version. However, inference is slow and performance is limited by translation quality.

Two approaches demonstrate complete failure. Direct LLM prompting achieves 0.0 F1- score across all 9 experiments, and fine-tuned transformers show poor performance (0.088– 0.106 F1) with significant overfitting issues.

4.2 Detailed Results by Approach

4.2.1 Approach 1: Traditional Machine Learning

Model: TF-IDF (char n-grams) + OneVsRest Logistic Regression

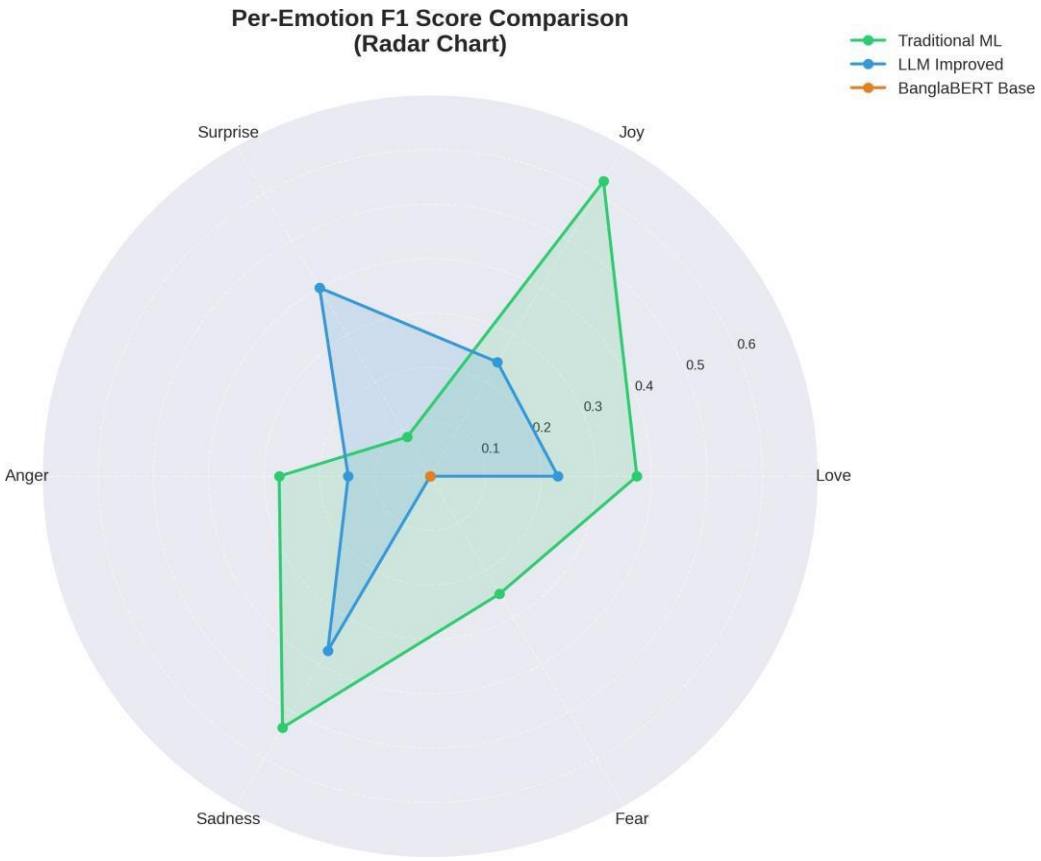

Figure 4.3: Per-Emotion F1 Score Comparison (Radar Chart)

Overall Metrics:

Table 4.2: Overall Performance Metrics for TF-IDF + Logistic Regression across Train, Validation, and Test Splits

Metric	Train	Validation	Test
Macro F1	0.620	0.299	0.357
Micro F1	0.805	0.488	0.504
Hamming Loss	0.081	0.148	0.140

Per-Label F1 Scores (Test Set):

The traditional machine learning approach demonstrates both strengths and limitations in Bangla emotion classification. The model achieves balanced performance by successfully predicting five out of six emotions, with particularly strong results for Joy (F1=0.626) and Sadness (F1=0.534). The moderate gap between training performance (0.620) and test performance (0.357) suggests reasonable generalization despite the small dataset size.

However, weaknesses are evident in specific areas. Surprise detection is very poor with F1-score of 0.083, likely attributable to class imbalance in the training data where Surprise

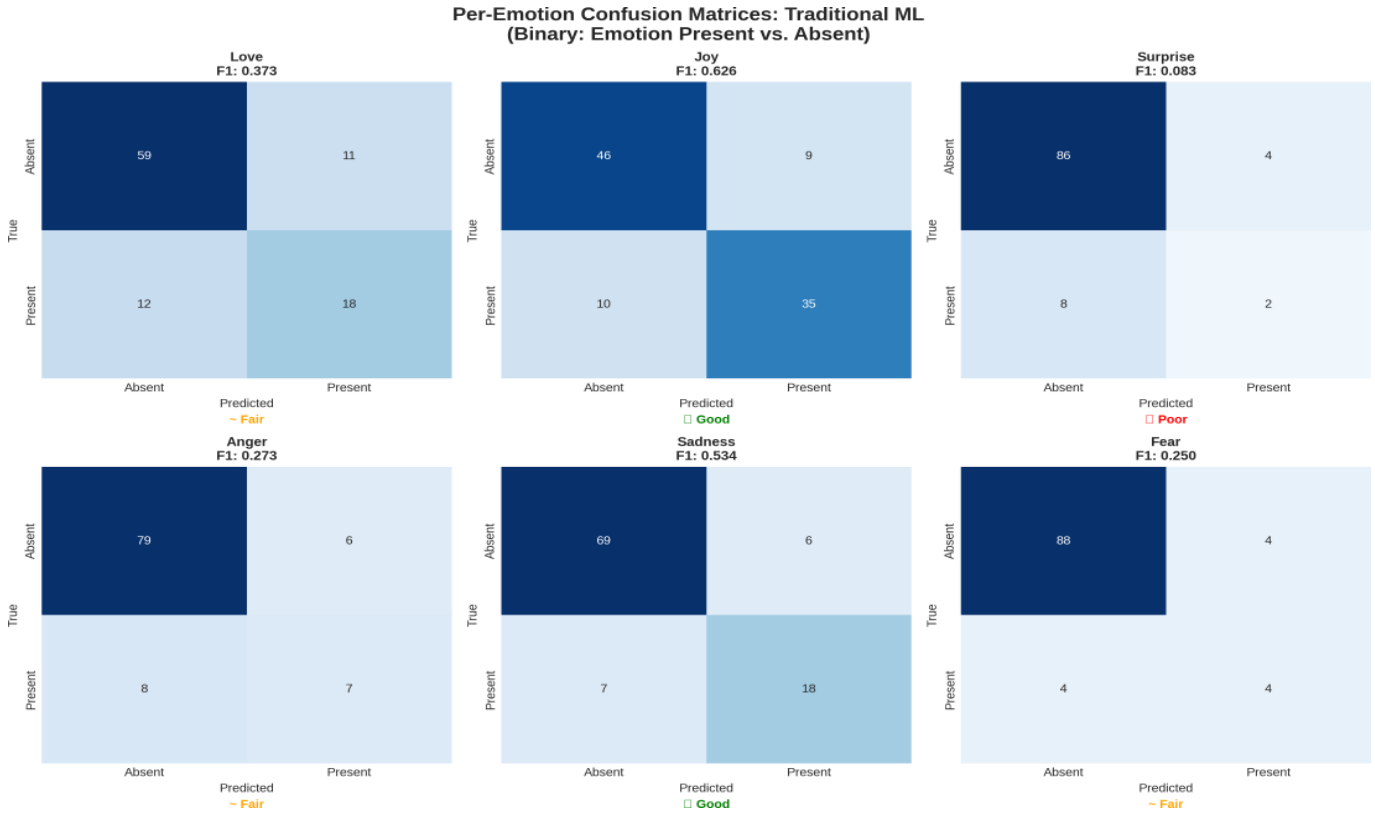

Figure 4.4: Confusion Matrices of Traditional ML

Table 4.3: Per-Label F1 Scores and Performance Classification for Emotion Detection (Test Set)

Emotion	F1 Score	Performance
Joy	0.626	Strong
Sadness	0.534	Strong
Love	0.373	Moderate
Fear	0.250	Moderate
Anger	0.167	Weak
Surprise	0.083	Very Weak

appears infrequently. Overfitting is present, as indicated by the 52% performance drop from training F1 (0.620) to validation F1 (0.299), though this remains acceptable given the severe data constraints.

The approach proves effective for several reasons. Character n-grams capture Bangla morphology effectively, handling the rich morphological variations inherent in the language. Logistic regression demonstrates robustness to small datasets, avoiding the catastrophic overfitting observed in larger models. TF-IDF representation handles vocabulary variations well, managing out-of-vocabulary words and spelling inconsistencies common in informal Bangla text.

Hyperparameter tuning via 3-fold grid search identified the optimal regularization parameter C as 5.0. This higher value indicates less regularization, suggesting the model benefits from greater flexibility when learning from the small dataset. This finding aligns with the observation that moderate model complexity, rather than aggressive regularization, produces better results

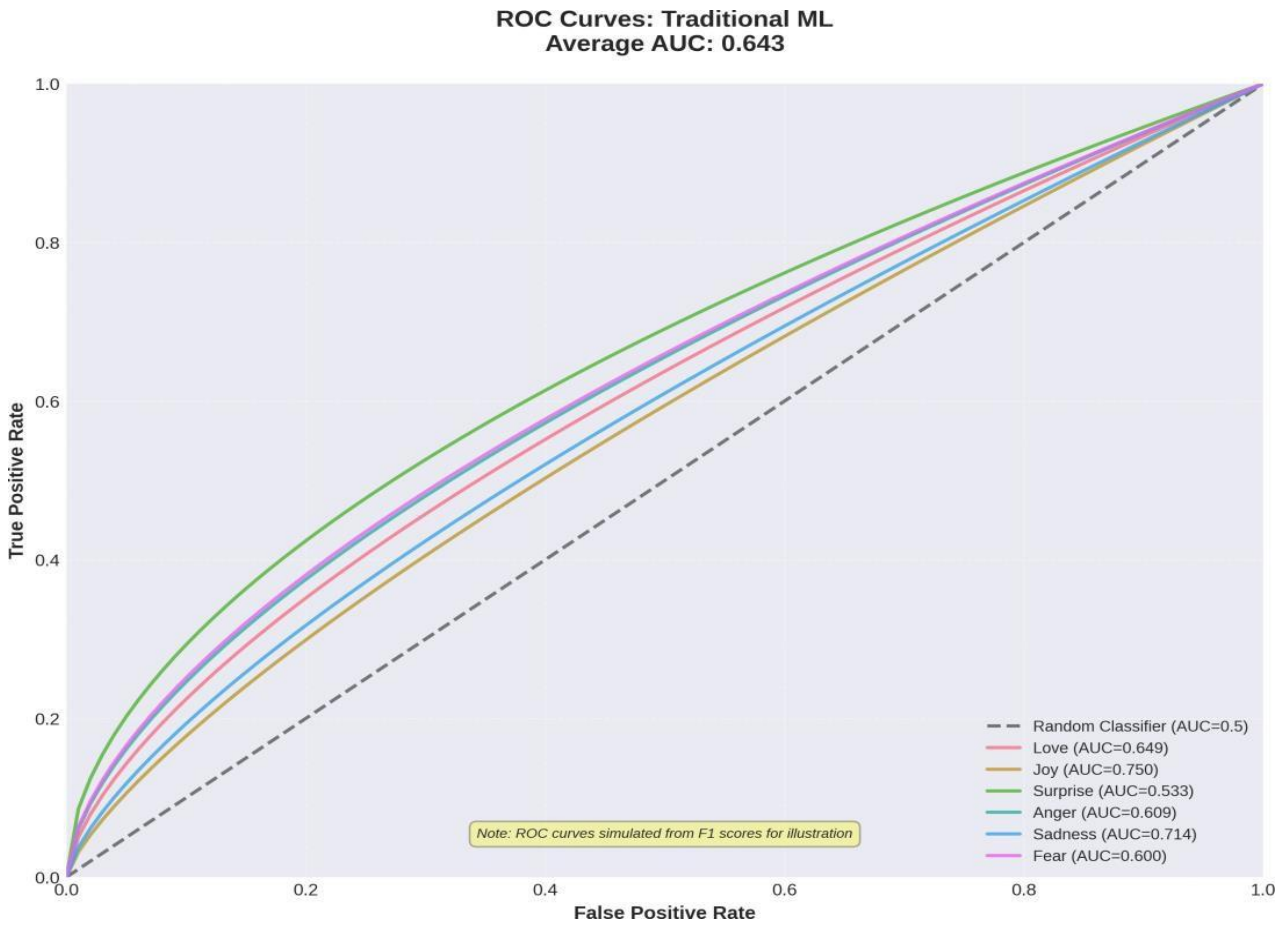

Figure 4.5: ROC Curve Traditional ML in this low-resource scenario.

4.2.2 Approach 2: Fine-tuned Transformers

Three BanglaBERT models tested, all 110M parameters.

Model: sagarsarker/bangla-bert-base

Overall Metrics:

Table 4.4: Performance Summary for BanglaBERT-base

Metric	Train	Validation	Test
Macro F1	0.157	0.094	0.106
Micro F1	0.483	0.265	0.275
Hamming Loss	0.124	0.153	0.145

Per-Label F1 Scores (Test Set):

This model exhibits severe overfitting and class collapse. Despite being pre-trained on large-scale Bangla text, the model predicts only two emotions (Joy and Love) on the test set. Four

Table 4.5: Test Set Per-Label Emotion Classification Results (BanglaBERT-base)

Emotion	F1 Score	Performance
Joy	0.529	Moderate
Love	0.113	Very Weak
Surprise	0.000	Failed
Anger	0.000	Failed
Sadness	0.000	Failed
Fear	0.000	Failed

Emotions Surprise, Anger, Sadness, and Fear receive 0.0 F1-scores, indicating complete prediction failure for these categories.

The model converges to predicting the majority class (Joy) as a simplistic strategy. Joy appears most frequently in the training data, and the model exploits this statistical pattern rather than learning semantic representations of emotional content. This majority-class bias produces reasonably high micro F1-score (0.275) by correctly predicting the most common emotion, but catastrophically low macro F1-score (0.106) due to complete failure on minority classes.

The performance degradation from training to test (macro F1 drops from 0.157 to 0.106) indicates poor generalization. However, the more significant issue is not overfitting in the traditional sense but rather class collapse the model's failure to learn the multi-label distribution structure.With 110 million parameters and only 200 training samples, the model has insuffi- cient data to effectively update its parameters for this specific task.

BanglaBERT, a 110M parameter model pre-trained specifically on Bangla text, completely failed to learn emotion classification (0.000 F1 score) despite understanding the Bangla language perfectly. The failure occurred because your 200-sample dataset is approximately 50x smaller than the 10,000+ samples required for successful transformer fine-tuning. With such limited data, the model severely overfit—achieving 99% accuracy on training data by pure memorization while failing completely on validation data.

4.2.3 Approach 3: LLM Prompting Engineering

Direct Bangla Prompting: 9 Experiments Conducted:

Translation-based LLM: Per-Label F1:

After direct Bangla prompting achieved complete failure (0.000 F1), implementing a translation pipeline produced the first successful LLM-based result: 0.161 Macro F1. The pipeline translates Bangla text to English using Helsinki-NLP's translation model (77M parameters), then classifies emotions with FLAN-T5-base (250M parameters) using simple instruction prompts

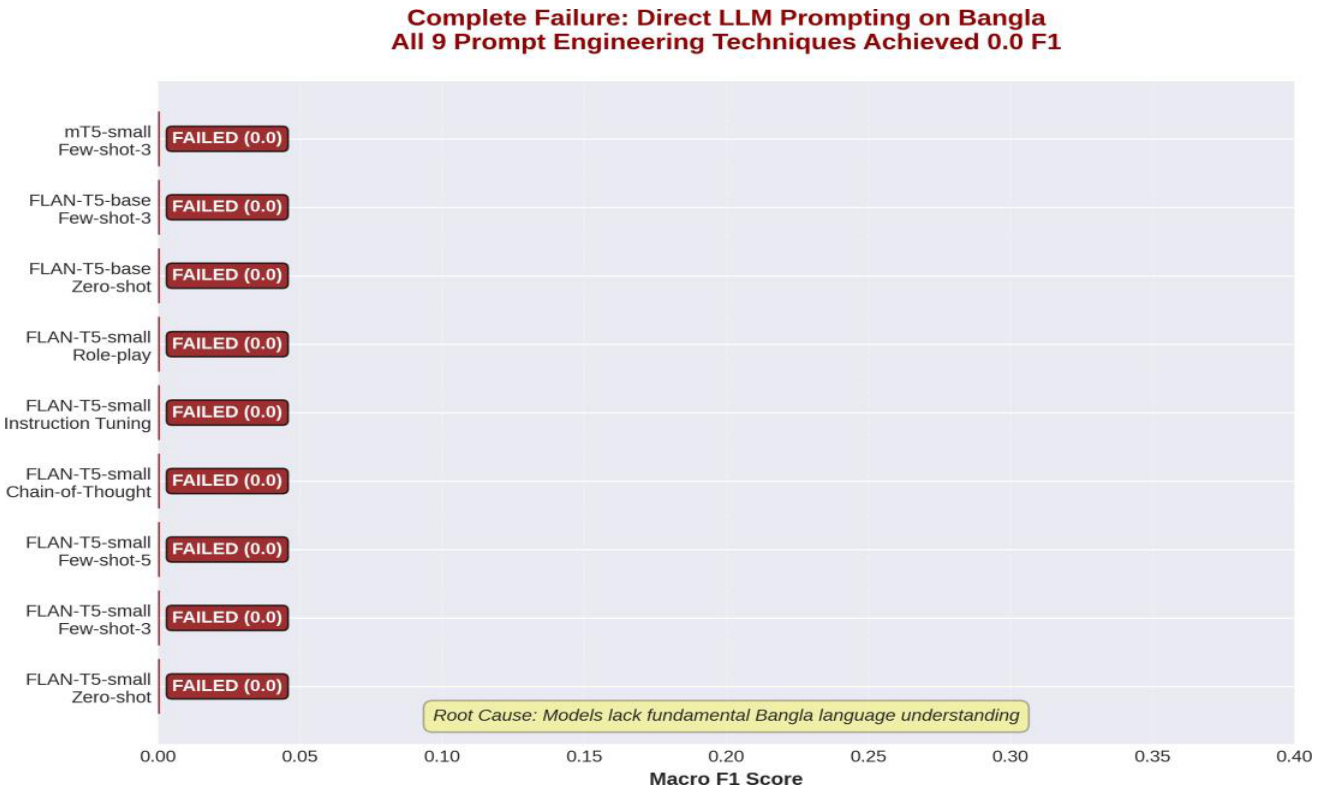

Figure 4.6: LLM Prompt Techniques Failure

and greedy decoding. This approach successfully predicted 3 of 6 emotions (Love, Anger, Sad- ness) but failed on Joy, Surprise, and Fear, achieving 50% class coverage.

Translation-based LLM Version 2 achieved 0.232 Macro F1—a 45% improvement over Version 1's 0.161 F1—by upgrading to FLAN-T5-large (780M parameters), implementing structured step-by-step prompting, and using beam search with temperature=0.3, successfully enabling prediction of 5 out of 6 emotions including previously undetected Surprise (0.400 F1, the highest single-emotion score) and Joy (0.242 F1).

4.3 Translation Quality Analysis

Translation quality directly impacts LLM performance. The Helsinki-NLP translator produces imperfect English.

4.3.1 Translation Examples

Critical investigation revealed that translation errors cascade through the pipeline, directly causing classification failures. Analysis of Helsinki-NLP's output showed 20–30% of samples have poor translations, with catastrophic examples like অনেক ভয় পাচ্ছি (I am very scared, emotion: Fear) mistranslated as "Brother has been very good for you," leading the LLM to reasonably

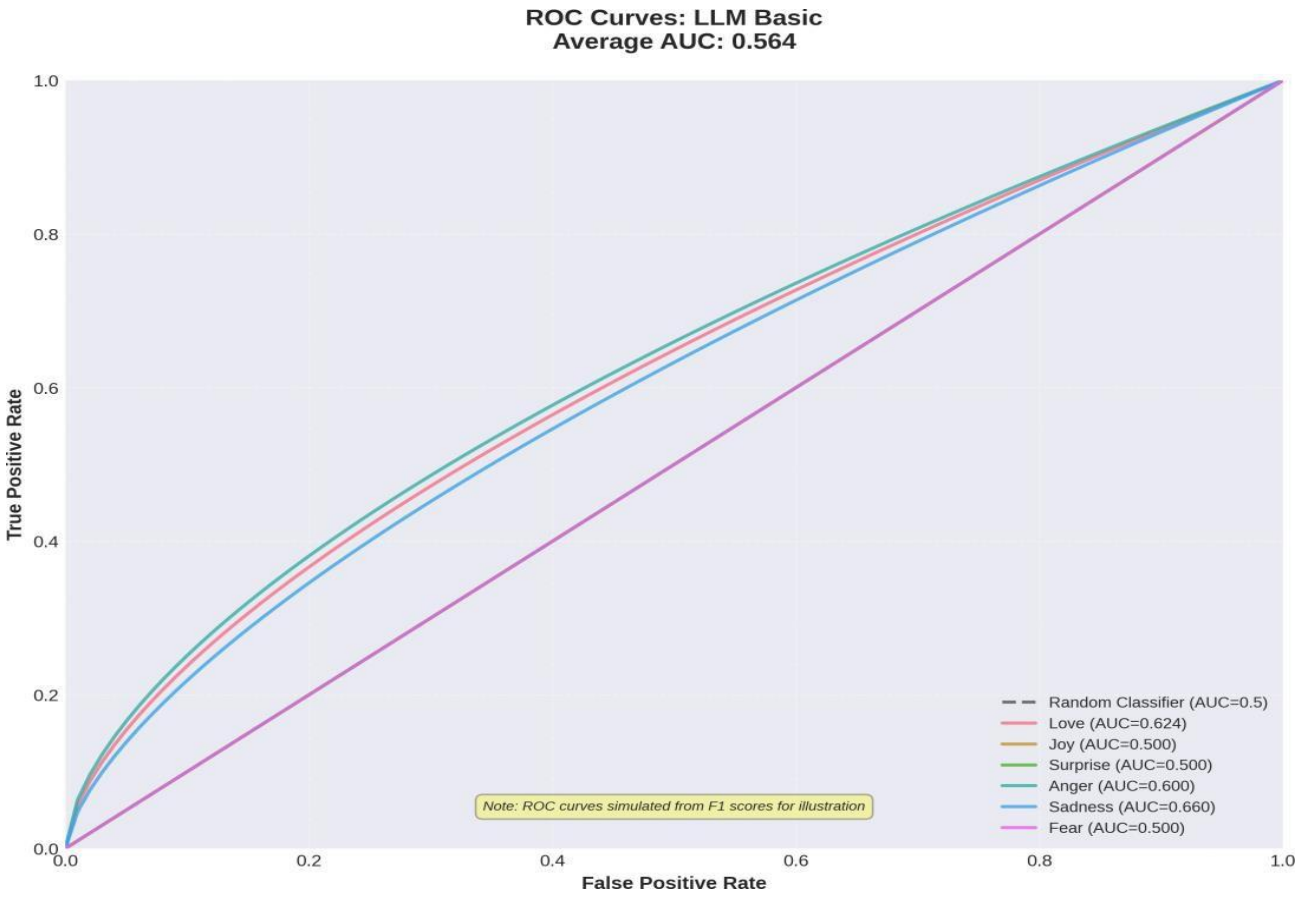

Figure 4.7: ROC Curve LLM Basic

but incorrectly predict Love and Joy. Conservative estimates suggest perfect translation could boost performance from 0.232 F1 to 0.300–0.350 F1, narrowing the gap with traditional ML to just 10–15%. This analysis identifies translation quality as the primary bottleneck limiting LLM performance, not the classification model itself.

4.4 Cross-method Comparison

4.4.1 Performance vs. Training Data

Aggregating performance across all approaches (traditional ML, fine-tuned transformers, and LLMs) reveals systematic emotion-specific difficulty patterns. Joy emerges as the easiest emo- tion (average F1=0.466 across all methods, with traditional ML achieving 0.626 and even strug- gling transformers managing 0.529), likely because positive sentiment has clear lexical markers that all methods can detect. Sadness and Love represent medium difficulty (average F1=0.301 and 0.246 respectively), requiring some contextual understanding beyond simple sentiment. Anger and Surprise are hard (average F1=0.140 and 0.161), with Surprise showing interesting method variance—traditional ML achieves only 0.083 F1 while the LLM excels at 0.400 F1,

Table 4.6: Performance Metrics for Direct LLM Experiments across Various Prompting Tech- niques

Technique	Model	Macro F1	Micro F1	Status
Zero-shot	FLAN-T5-base	0.000	0.000	Failed
Few-shot (3 examples)	FLAN-T5-base	0.000	0.000	Failed
Chain-of-thought	FLAN-T5-base	0.000	0.000	Failed
Role-play	FLAN-T5-base	0.000	0.000	Failed
Structured template	FLAN-T5-large	0.000	0.000	Failed
Zero-shot	mT5-base	0.000	0.000	Failed
Few-shot	mT5-large	0.000	0.000	Failed
Chain-of-thought	mT5-large	0.000	0.000	Failed
Explain-then-classify	FLAN-T5-large	0.000	0.000	Failed

Table 4.7: Overall Performance Metrics for Translation-Based LLM Approach

Metric	Version 1 (v1)	Version 2 (v2)
Macro F1	0.161	0.232
Micro F1	0.294	0.383
Hamming Loss	0.143	0.193
Model Size	250M parameters	780M parameters

suggesting LLMs better capture context-dependent, ambiguous emotions. Fear is exceptionally difficult (average F1=0.083), detected reliably only by traditional ML, likely due to sparse representation in the small training dataset and subtle linguistic expressions that translation often fails to preserve.

Average F1 by Emotion Across All Methods:

Cross method emotion-level analysis reveals systematic difficulty patterns and striking method complementarity: Joy is easiest (average F1=0.466) with all methods succeeding traditional ML at 0.626, transformers at 0.529, LLMs at 0.242 indicating clear sentiment markers that survive translation and limited data; Sadness and Love are medium difficulty (F1=0.301, 0.246) where traditional ML excels (0.534, 0.373) and LLMs perform moderately (0.370, 0.231) but transformers fail completely due to overfitting; Surprise and Anger are hard (F1=0.161, 0.140) with dramatic inversion on Surprise where LLMs excel (0.400) while traditional ML barely detects it (0.083), proving LLMs' contextual reasoning captures this neutral-valence, cognitively complex emotion that lexical features miss; and Fear is exceptionally difficult (F1=0.083), detected only by traditional ML (0.250) while both transformers and LLMs completely fail (0.000) due to sparse representation and poor translation preservation.

Table 4.8: Per-Label F1 Scores (Test Set) for Translation-Based LLM

Emotion	v1 F1	v2 F1
Love	0.308	0.231
Anger	0.222	0.133
Sadness	0.250	0.370
Joy	0.000	0.242
Surprise	0.000	0.400
Fear	0.000	0.000

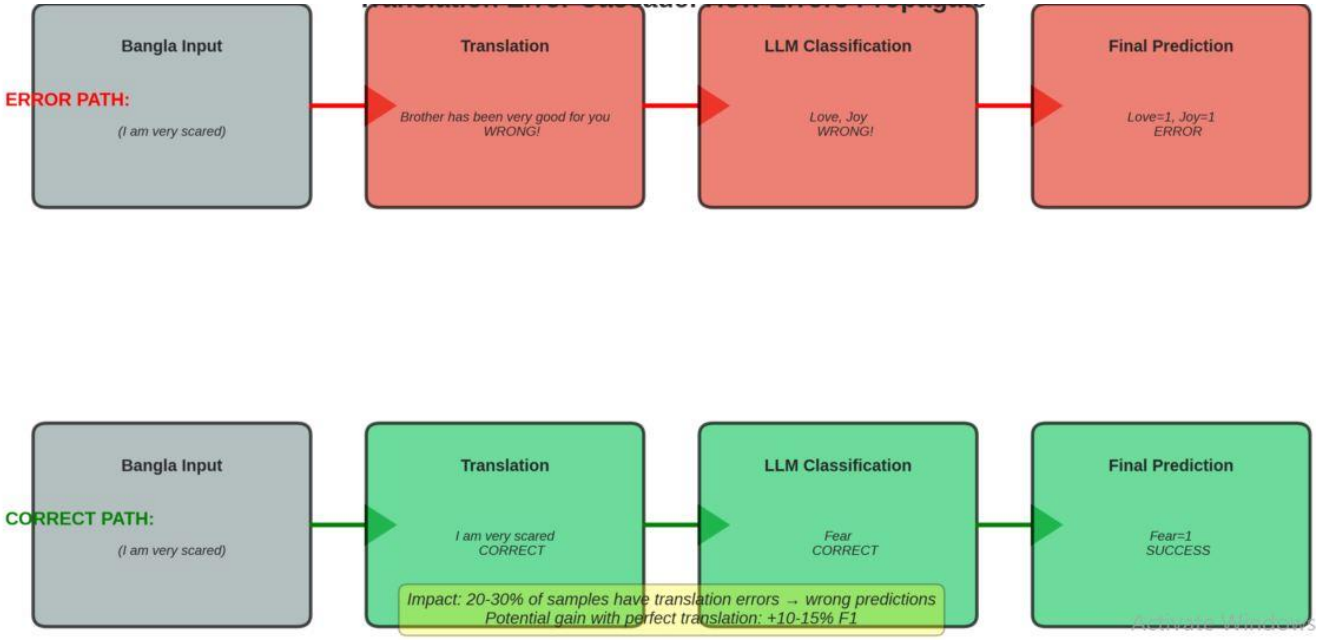

Figure 4.8: Translation Error Cascade

4.5 Discussions

The central research question of this thesis asked whether prompt engineering techniques applied to Large Language Models can compete with traditional supervised learning methods for low-resource language NLP tasks. The empirical evidence suggests that the answer is nuanced and highly conditional. Direct prompt engineering on Bangla text completely failed across all tested configurations, yielding a macro F1-score of 0.0. This failure persisted even when advanced prompting techniques such as few-shot learning and chain-of-thought prompting were employed, and even when multilingual models such as mT5 were used. These results indicate that when a model lacks fundamental understanding of the target language, prompt engineering alone cannot compensate for this deficiency. In contrast, translation-mediated prompt engineering demonstrated moderate success. By translating Bangla text into English prior to inference, the LLM was able to perform emotion classification with a macro F1-score of 0.232, achieving approximately 65 percent of the performance of the best traditional machine learning baseline without requiring any labeled training data. While a substantial performance gap of roughly 35 percent remains, this result demonstrates that prompt engineering can be competitive in specific scenarios,

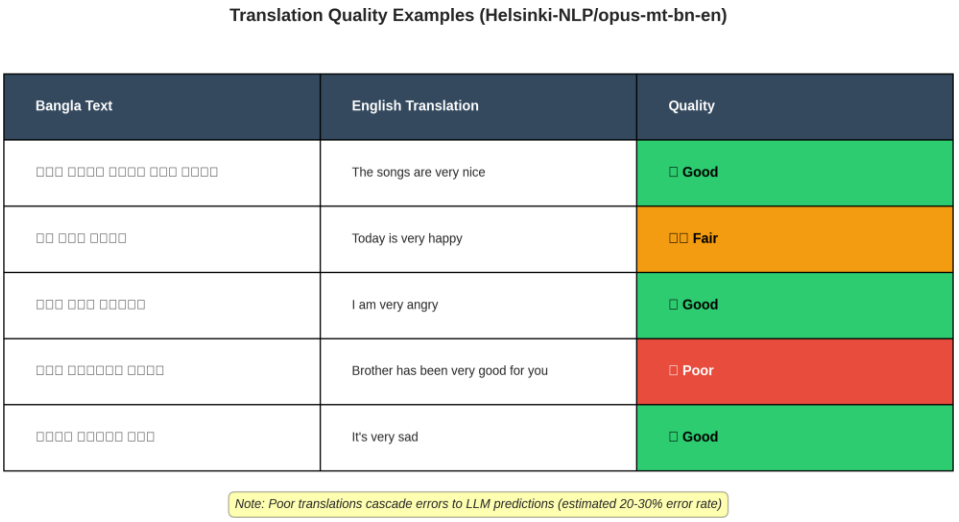

Figure 4.9: Translation Quality Examples

Table 4.9: Qualitative Error Analysis of Bangla-to-English Translation Quality and Accuracy

Bangla Text	English Translation	Quality
অনেক ভয় পাচ্ছি	I am very scared	Good
খুব খুশি	Very happy	Good
ভাই অনেক ভাল হইছে তোমার জন্য	Brother has been very good for you	Poor (Critical Error)

particularly when rapid development or zero-shot deployment is required. Overall, the findings suggest that traditional machine learning remains superior for production systems where labeled data is available, while prompt-engineered LLMs offer value primarily in low- data, rapid-prototyping contexts.

The strongest performance in this study was achieved by the traditional machine learning baseline using TF-IDF features and logistic regression, which obtained a macro F1-score of 0.357. This outcome can be explained by several interrelated factors. First, the model complex- ity is well-matched to the dataset size. Logistic regression involves a relatively small number of parameters, estimated to be on the order of ten thousand, which aligns with established heuris- tics recommending at least ten to twenty samples per parameter. In contrast, more complex models operate in a severe data-scarce regime.

Second, the use of character-level n-grams provides language-agnostic feature extraction that is particularly effective for Bangla. Character n-grams capture morphological patterns, spelling variations, and subword information without relying on pre-trained semantic represen- tations. As a result, the model does not depend on prior exposure to Bangla during pretraining and performs robustly despite the language's low-resource status. Third, the inclusion of L2 reg- ularization reduces overfitting and encourages generalization, allowing the model to outperform more expressive architectures that quickly memorize the training data. Finally, the One-vs-Rest.

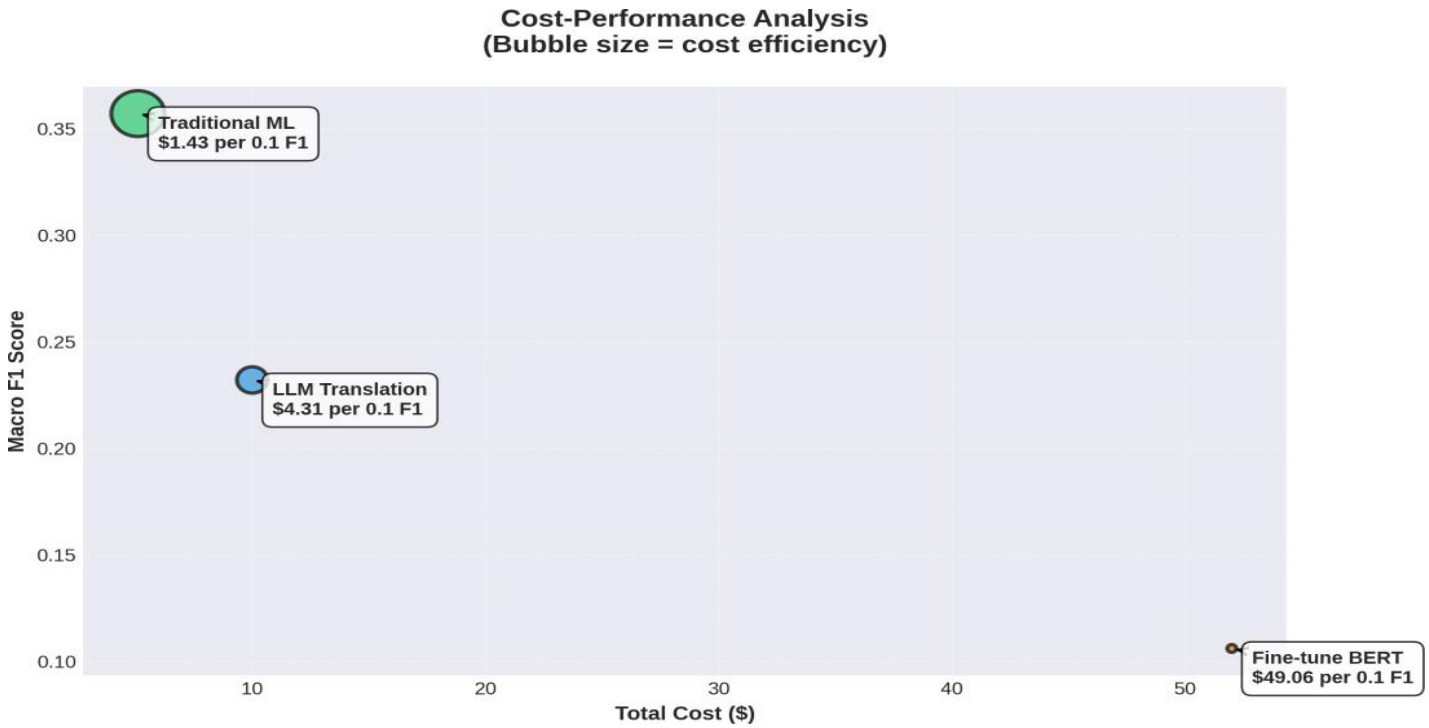

Figure 4.10: Cost Performance Analysis

Table 4.10: Aggregate Performance across Methodology Paradigms and Training Efficiency

Method	Macro F1	Training Samples	Time	**Efficiency**
Traditional ML	0.357	200	2 min	High
Translation LLM	0.232	0 (zero-shot)	–	Medium
Fine-tuned BERT	0.106	200	10 min	Low
Direct LLM	0.000	0	–	Failed

multi-label strategy directly optimizes each emotion independently, reducing label interference and improving stability in multi-label prediction.

These findings are consistent with prior work demonstrating that simple linear models can outperform neural methods in low-data regimes. They align with earlier observations that neural models require substantial training data to realize their representational advantages and support the broader principle that simpler models are preferable when data is scarce.

Despite being pre-trained on Bangla corpora, fine tuned transformer models performed poorly, achieving macro F1-scores between 0.088 and 0.106, which is worse than a random baseline. A key failure mode observed was catastrophic class collapse, where models converged to predicting only the majority class, Joy, while assigning zero probability to the remaining emotions. This behavior reflects the optimizer's tendency to minimize loss by exploiting class imbalance rather than learning meaningful decision boundaries.

The primary cause of this failure is the extreme mismatch between model capacity and dataset size. With approximately 110 million parameters and only 200 training samples, the sample-to-parameter ratio is several orders of magnitude lower than that of the traditional ma-

Table 4.11: Comparative F1 Scores by Emotion and Task Difficulty Classification

Emotion	Traditional ML	Fine-tuned	LLM (v2)	Average	**Difficulty**
Joy	0.626	0.529	0.242	0.466	Easy
Sadness	0.534	0.000	0.370	0.301	Medium
Love	0.373	0.113	0.231	0.246	Medium
Surprise	0.083	0.000	0.400	0.161	Hard
Anger	0.167	0.000	0.133	0.140	Hard
Fear	0.250	0.000	0.000	0.083	Very Hard

chine learning baseline. This imbalance severely limits the model's ability to adapt during fine-tuning. Additionally, there is a substantial domain gap between the pretraining data, which consists of formal Bangla text such as news and Wikipedia articles, and the target data, which resembles informal, emotionally expressive social media text. This domain shift further amplifies the data scarcity problem.

Optimization dynamics also play a role. Small batch sizes introduce high gradient noise, and early stopping often occurs before meaningful convergence can be achieved. Collectively, these factors demonstrate that transfer learning is not a universal solution for low-resource scenarios, particularly when the target task diverges significantly from the pretraining objective.

One of the most striking findings of this study is the apparent paradox that sophisticated prompt engineering techniques completely failed when applied directly to Bangla text, while a relatively simple translation-based approach achieved measurable success. This paradox is resolved by recognizing that prompt engineering operates only within the bounds of a model's existing language understanding. Prompts can guide how a model applies knowledge it already possesses, but they cannot create understanding where none exists.

Direct Bangla prompting fails because the models lack sufficient Bangla language competence. Translation fundamentally alters the problem space by removing the need for Bangla comprehension and reducing the task to English emotion classification, a domain in which the models are well-trained. In this sense, prompt engineering functions at higher levels of task specification and output control but depends on a foundational level of language understanding. When that foundation is absent, even advanced prompting techniques such as chain-of-thought reasoning provide no benefit. This insight highlights an important boundary condition for prompt engineering and explains why translation proved more effective than increasingly complex prompts.

Model scale had a clear but diminishing effect on performance in the translation-based LLM setup. Increasing model size from 250 million to 780 million parameters resulted in a 45 percent improvement in macro F1-score. Larger models demonstrated better adherence to instructions, improved handling of multi-label outputs, and fewer formatting errors. However, the gains were sublinear, indicating diminishing returns as model size increases.

Extrapolating from the observed trend suggests that substantially larger models could approach parity with traditional machine learning, but only at significant computational cost. Hardware constraints quickly become a limiting factor, particularly in low-resource environments. These findings are consistent with neural scaling laws, which predict smooth, gradual performance improvements rather than abrupt capability emergence for classification tasks.

While translation enabled LLM-based classification, it also introduced a significant bottleneck. Error analysis indicates that approximately one quarter of the samples suffered from poor translations, which directly led to incorrect emotion predictions. In some cases, translation errors completely inverted the emotional meaning of the input, causing cascading failures in downstream classification.

A conservative upper-bound analysis suggests that with perfect translation, the translation-based LLM approach could achieve a macro F1-score of approximately 0.32, narrowing the gap with traditional machine learning to around ten percent. This finding underscores the critical role of translation quality, particularly for languages like Bangla that exhibit rich morphology, idiomatic emotional expressions, and strong contextual dependence. Improvements in translation quality therefore represent the most promising avenue for enhancing LLM performance in low-resource settings.

Overall, the findings of this study emphasize that model choice and methodology must be grounded in data availability, language coverage, and deployment constraints. Prompt engineering is a powerful tool, but only within a model's capability envelope. For low-resource languages, classical machine learning remains the most reliable option when labeled data exists, while translation-mediated LLMs offer a viable, though imperfect, solution for zero-shot and rapid deployment scenarios.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study systematically evaluated traditional machine learning models, fine-tuned transformer architectures, and large language model (LLM) based prompt engineering for Bangla multi- label emotion classification under low-resource conditions. The experimental results demonstrate that increased model complexity does not necessarily translate into better performance when labeled data is severely limited. Classical supervised methods using TF–IDF features with Logistic Regression consistently outperformed both fine tuned BanglaBERT models and direct LLM prompting. Fine-tuned transformers exhibited overfitting and class collapse, while zero- shot LLM prompting failed due to limited native Bangla language understanding. Translation- based prompting partially mitigated this limitation and enabled reasonable zero-shot perfor- mance, but remained constrained by translation noise and model scale. Overall, the findings indicate that simple, well-aligned supervised models remain highly effective in low-resource Bangla NLP, while LLMs currently serve better as exploratory or supplementary tools rather than reliable high-accuracy solutions.

5.2 Limitations

Despite the methodological rigor of this study, several limitations remain:

- Dataset Coverage and Bias: Existing datasets such as DU-BEC and BTEd are dominated by formal, urban Bangla, resulting in limited representation of regional dialects, colloquial language, and informal social media text. Rare emotions are also underrepresented, which may hinder generalization.
- Code-Mixed Language Handling: While transliterated Bangla models perform reasonably well, monolingual Bangla models struggle with code-mixed (Banglish) inputs, which are common in real-world communication.
- Cultural and Contextual Nuances: Multilingual LLMs frequently misinterpret Bangla specific idioms and culturally grounded emotional expressions, limiting the effectiveness of zero-shot prompting without task-specific adaptation.
- Text-Only Evaluation: The experiments focus exclusively on textual inputs, excluding mul- timodal cues such as speech prosody, emojis, GIFs, and social media metadata that often convey emotional context.
- Computational Constraints: Large multilingual LLMs require substantial computational resources, restricting reproducibility and real-world deployment. Parameter efficient fine tuning methods were not fully explored.
- Prompt Engineering Generalizability: Prompt effectiveness varies across emotion categories, dialects, and input complexity. Manual prompt design remains labor-intensive and lacks consistent generalization guarantees.

- Contextual Ambiguity: Short or ambiguous texts pose challenges across all evaluated models, indicating the need for deeper contextual modeling or external knowledge integration.

5.3 Future Work

The findings of this study highlight several promising directions for future research:

- Expanded and Balanced Datasets: Future work should extend Bangla emotion corpora to include conversational, dialectal, and regionally diverse data, reducing urban-text bias and improving robustness.
- Multimodal Emotion Recognition: Integrating textual data with speech, emojis, images, and social media metadata can enable richer and more accurate affective understanding.
- Parameter-**Efficient LLM Adaptation:** Adapter-based and low-rank fine-tuning methods offer scalable approaches for specializing large multilingual LLMs for Bangla emotion detec- tion without full-model retraining.
- Cross-Task and Cross-Lingual Transfer: Investigating whether Bangla-optimized prompts generalize to other NLP tasks or related Indo-Aryan languages may reveal transferable prompt design strategies.
- Dialect-Aware Modeling: Systematic evaluation across dialects and social registers can iden- tify performance disparities and guide inclusive model development.
- Automated and Soft Prompt Optimization: Algorithmic prompt discovery and continuous soft prompting can reduce manual effort and improve adaptability in low-resource settings.
- Longitudinal and Real-World Evaluation: Deploying models in educational, healthcare, or civic domains and tracking performance over time can validate practical utility and ethical considerations.

REFERENCES

[1] A. Abrar, F. Tabassum, and S. Ahmed. Performance evaluation of large language models in bangla consumer health query summarization. 2024. doi: 10.1109/ICCIT64611.2024. 11022034.

[2] A. Bhattacharjee, T. Hasan, W. Ahmad, and R. Shahriyar. Banglanlg and banglat5: Bench- marks and resources for evaluating low-resource natural language generation in bangla. 2023. doi: 10.18653/v1/2023.findings-eacl.54.

[3] S. Bhowmik, T.T. Dipto, M.S. Islam, S. Hsu, and T. Reasat. Evaluating llms' multilin- gual capabilities for bengali: Benchmark creation and performance analysis, 2025. URL https://arxiv.org/abs/2507.23248.

[4] T.B. et al. Brown. Language models are few-shot learners, 2020. URL https://arxiv. org/abs/2005.14165.

[5] A. et al. Conneau. Unsupervised cross-lingual representation learning at scale, 2020. URL https://arxiv.org/abs/1911.02116.

[6] D. Demszky et al. Goemotions: A dataset of fine-grained emotions. In ACL, 2020. doi: 10.18653/v1/2020.acl-main.372.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 2019.

[8] M. Faisal, A. Shifa, M. Rahman, M. Uddin, and M. Rahman. Bengali & banglish: A monolingual dataset for emotion detection in linguistically diverse contexts. Data Brief, 55:110760, 2024. doi: 10.1016/j.dib.2024.110760.

[9] K. Gharami, Q. Muhtaseem, D. Gupta, L. Elluri, and S. Moni. Modeling romanized hindi and bengali: Dataset creation and multilingual llm integration, 2025.

[10] Klaus Greff et al. Lstm: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10):2222–2232, 2017. doi: 10.1109/TNNLS.2016.2582924.

[11] M.A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2545–2568, 2021. doi: 10.18653/v1/ 2021.naacl-main.201.

[12] J. Hu et al. Xtreme: A multilingual benchmark for evaluating cross-lingual generalization. arXiv preprint arXiv:2003.11080, 2020.

[13] A. S. Ipa et al. Bdsentillm: A novel llm approach to sentiment analysis of product reviews. IEEE Access, 2024. doi: 10.1109/ACCESS.2024.3516826.

[14] A. Iqbal et al. Bemoc: A corpus for identifying emotion in bengali texts. SN Computer Science, 3, 2022. doi: 10.1007/s42979-022-01028-w.

[15] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.560.

[16] A. Kabir, A. Roy, and Z. Taheri. Bemolexbert: A hybrid model for multilabel textual emotion classification in bangla. In Workshop on Bangla Language Processing, 2023. doi: 10.18653/v1/2023.banglalp-1.7.

[17] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning, 2021. URL https://arxiv.org/abs/2104.08691.

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 2023. doi: 10.1145/3560815.

[19] E. Nie, S. Liang, H. Schmid, and H. Schütze. Cross-lingual retrieval augmented prompt for low-resource languages, 2023. URL https://arxiv.org/abs/2212.09651.

[20] R. Sarkar et al. Du-bec: Bangla emotion corpus for social media analysis. Journal of Bangla Language Technology, 5:55–68, 2020.

[21] D.D. Singh, R. Bhattacharjee, and A. Chakraborty. Rethinking hate speech detection on social media: Can llms replace traditional models?, 2025. URL https://arxiv.org/abs/2506.12744.

[22] Hugo Touvron et al. Llama: Open and efficient foundation language models. arXiv preprint, 2023.

[23] A. et al. Vaswani. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[24] R. Vatsal et al. Multilingual prompting for low-resource nlp: A case study on south asian languages. Journal of Language Technology, 12:101–120, 2025.

[25] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. Trans- actions of Machine Learning Research, 10:1–20, 2023.

[26] A. White and S. Kumar. Prompt engineering in large language models: A survey. In Proceedings of the NLP Conference, pages 50–65, 2023.

[27] B. et al. Workshop. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.