CSE366 Project Report

Emotion Detection in Bangla with LLM

Junnun Mohamed Karim (2022-1-60-108)
Md Murad Khan Limon (2022-1-60-044)
Md. Yousuf Hozaifa (2022-1-60-162)

May 31, 2024

Abstract

This work investigates the application of Large Language Models (LLMs) for emotion detection in Bangla text. We fine-tune the Gemma-2B model on a dataset annotated with five emotions: Happy, Love, Sadness, Anger, and Fear. Our methodology tackles class imbalance through stratified splitting and data preprocessing. The fine-tuned model achieves an overall accuracy of 71.2% on the test set. However, performance varies across emotions, with the model excelling at identifying Happy emotions (F1-score: 0.81) but struggling with underrepresented classes like Fear (F1-score: 0.00). This study highlights the potential of LLMs for Bangla emotion analysis while revealing the challenges posed by class imbalance. Future work explores data augmentation and advanced techniques to improve performance on minority classes.

Contents

1 Introduction

Emotion detection is crucial for various applications such as sentiment analysis, human-computer interaction, and mental health monitoring. Despite the advances in NLP for widely spoken languages, resources for Bangla are limited. This project leverages a Large Language Model (LLM) to bridge this gap, aiming to accurately detect emotions in Bangla text. The project objectives include dataset preparation, model fine-tuning, quantization for efficient deployment, and developing a user-friendly interface using Django.

2 Methodology

This section outlines the detailed methodology followed in this project for fine-tuning the Gemma-2B model to perform emotion detection in Bangla. The methodology is divided into several sub-sections, covering data acquisition, preprocessing, model preparation, training, and evaluation.

2.1 Data Acquisition

The dataset used for this project is sourced from the paper titled *"A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews"* [Ras+24]. The dataset comprises Bangla sentences annotated with emotion labels. This dataset includes reviews from various e-commerce platforms in Bangladesh, categorized into five distinct emotions: Happy, Love, Sadness, Anger, and Fear.

2.2 Dataset Imbalance

Upon examining the dataset, it was evident that the classes were imbalanced. The distribution of the emotion labels was as follows:

- Happy: 41,971 instances
- Love: 18,570 instances
- Sadness: 5,880 instances
- Anger: 2,854 instances
- Fear: 1,042 instances

This imbalance poses a challenge for training the model, as it may lead to a bias towards the more frequent classes (Happy and Love) and underperformance on the less frequent ones (Sadness, Anger, and Fear). Addressing this imbalance is crucial for developing a robust emotion detection model.

2.3 Data Preprocessing

The data preprocessing steps involved loading the dataset and performing several transformations to prepare it for model training:

1. **Loading the Dataset:** The dataset was loaded using the `pandas` library, allowing us to analyze the data structure and perform initial inspections.
2. **Class Distribution Analysis:** We visualized the distribution of emotion classes to understand the extent of imbalance and to plan strategies to mitigate its impact.
3. **Stratified Splitting:** To ensure each split of the dataset maintains the same class distribution, we performed stratified splitting, dividing the dataset into training, validation, and test sets.
4. **Prompt Generation:** We created formatted prompts for the model training. For each review, a prompt was generated in the form: `"Analyze the emotion of the text enclosed in square brackets, determine if it is Happy, Love, Sadness, Anger or Fear, and return the answer only as the corresponding sentiment label 'Happy' or 'Love' or 'Sadness' or 'Anger' or 'Fear'."`

2.4 Model Preparation

We employed the Gemma-2B model from Google's Hugging Face repository. To fit the model within memory constraints and optimize its performance on the specific task, the following configurations were applied:

- **Maximum Sequence Length:** Set to 2048 tokens to handle long sequences.
- **4-bit Quantization:** Enabled to reduce memory usage while maintaining performance.
- **Parameter-Efficient Fine-Tuning (PEFT):** Applied using LoRA (Low-Rank Adaptation) to fine-tune the model with fewer parameters, focusing on specific modules such as query, key, value, and output projections.

2.5 Training

The training process involved several steps to fine-tune the model effectively:

1. **Training Configuration:** Set up training arguments, including batch size, learning rate, number of steps, and optimizer type. We used the AdamW optimizer with 8-bit precision to handle the large-scale model efficiently.
2. **Trainer Initialization:** Utilized the `SFTTrainer` from the `trl` library to handle the fine-tuning process, specifying the training and validation datasets, prompt field, and other configurations.
3. **Training Execution:** Conducted the training for 1800 steps, logging metrics such as loss at regular intervals.

2.6 Evaluation and Inference

Post-training, the model's performance was evaluated using standard metrics:

1. **Accuracy and Classification Report:** Calculated accuracy and generated a classification report, detailing precision, recall, and F1-score for each emotion class.
2. **Confusion Matrix:** Generated to visualize the performance across different classes and identify misclassifications.
3. **Prediction on Test Set:** The fine-tuned model was used to predict emotions on the test set, and the results were compared with actual labels to assess performance.

2.7 Model Quantization and Deployment

For efficient deployment on local machines, the model was quantized:

- **Quantization Method:** Applied `q4_k_m` quantization method, balancing performance and resource usage.
- **Conversion to GGUF:** Saved the quantized model in GGUF format for compatibility with `llama.cpp`.
- **Django Frontend Integration:** Developed a Django frontend to allow users to interact with the model locally, providing an interface for emotion detection on Bangla text inputs.

The comprehensive methodology ensured that the model was fine-tuned and optimized for the specific task of Bangla emotion detection, while addressing the challenges of class imbalance and efficient deployment.

3 Results

3.1 Dataset Statistics

The dataset used for fine-tuning the GEMMA-2B model consisted of 78,130 samples divided into training, testing, and evaluation sets. The distribution of the dataset is presented in Table 1. The training set contained 70,317 samples, the test set 3,906 samples, and the evaluation set 3,907 samples.

Emotion	**Count**	**Percentage**
Happy	41,971	53.7%
Love	18,570	23.8%
Sadness	5,880	7.5%
Anger	2,854	3.6%
Fear	1,042	1.3%
Total	78,130	100%

Table 1: Distribution of emotions in the dataset.

3.2 Model Performance

The model was evaluated using the test set, which consisted of 3,906 samples. The performance metrics, including accuracy, precision, recall, and F1-score, are summarized in Table 2. The overall accuracy of the model was 71.2%.

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Fear (1)	0.00	0.00	0.00	58
Anger (2)	0.73	0.52	0.61	159
Sadness (3)	0.55	0.66	0.60	326
Love (4)	0.68	0.40	0.50	1,031
Happy (5)	0.74	0.89	0.81	2,332
Accuracy				71.2%
Macro avg	0.54	0.49	0.50	3,906
Weighted avg	0.70	0.71	0.69	3,906

Table 2: Classification report of the fine-tuned GEMMA-2B model on the test set.

3.3 Confusion Matrix

The confusion matrix, shown in Table 3, provides a detailed breakdown of the model's predictions against the true labels for each emotion category. The matrix highlights that the model performed best in identifying the "Happy" emotion, with an accuracy of 89%. The model struggled with "Fear," achieving no correct predictions for this category.

	Fear	Anger	Sadness	Love	Happy
Fear (1)	0	6	32	0	20
Anger (2)	0	83	60	1	15
Sadness (3)	0	20	214	1	91
Love (4)	0	2	18	408	603
Happy (5)	0	2	66	189	2,075

Table 3: Confusion matrix of the fine-tuned GEMMA-2B model on the test set.

4 Discussion

The objective of this study was to fine-tune a large language model for the task of Bangla emotion analysis using the Gemma-2B model. The dataset used for training and evaluation comprised 70,317 training samples, 3,906 test samples, and 3,907 evaluation samples, each annotated with one of the following emotions: Happy, Love, Sadness, Anger, or Fear. The class distribution indicated a significant imbalance, with the majority of samples labeled as Happy (41,971) and the minority as Fear (1,042).

4.1 Performance Analysis

The fine-tuned model achieved an overall accuracy of 71.17% on the test set. Table 4 provides a detailed classification report, showing the precision, recall, and F1-score for each emotion category.

Emotion	Precision	Recall	F1-Score
Fear (1)	0.00	0.00	0.00
Anger (2)	0.73	0.52	0.61
Sadness (3)	0.55	0.66	0.60
Love (4)	0.68	0.40	0.50
Happy (5)	0.74	0.89	0.81

Table 4: Classification Report for Bangla Emotion Analysis

4.2 Challenges and Limitations

4.2.1 Class Imbalance

One of the major challenges observed was the significant class imbalance in the dataset. Emotions such as Fear and Anger had substantially fewer samples compared to Happy and Love. This imbalance is reflected in the performance metrics, particularly for the Fear class, where the model failed to make correct predictions (Precision, Recall, and F1-Score all being 0.00). This suggests that the model had insufficient data to learn from for the Fear class, resulting in poor generalization.

4.2.2 Recall and Precision Trade-offs

The precision and recall trade-offs varied significantly among different emotions. For instance, while the model achieved high recall for the Happy class (0.89), its recall for the Love class was relatively low (0.40). This indicates that while the model was good at identifying happy emotions, it struggled to correctly identify all instances of love, leading to more false negatives in that category.

4.3 Confusion Matrix Analysis

The confusion matrix in Table 3 provides further insights into where the model made errors.

The confusion matrix reveals several key patterns:

- The model frequently misclassified Fear and Anger as Sadness, indicating a potential overlap in the emotional expressions of these categories within the dataset.
- There was a notable confusion between Love and Happy, as many instances of Love were incorrectly classified as Happy. This suggests that these emotions might share linguistic similarities that the model finds challenging to distinguish.
- The model performed best at identifying Happy emotions, which aligns with the high recall and precision metrics for this class. However, some instances of Sadness were also misclassified as Happy, indicating that the model sometimes interprets negative emotions as positive.

4.4 Future Work

To address the limitations identified in this study, future work could focus on the following:

- **Data Augmentation:** Increasing the number of samples for underrepresented classes like Fear and Anger could help improve model performance for these categories.
- **Advanced Techniques:** Implementing more sophisticated techniques such as synthetic data generation, transfer learning from emotion recognition datasets in other languages, and using class weights during training to mitigate the effects of class imbalance.
- **Model Enhancement:** Exploring other model architectures or combining multiple models (ensemble methods) to enhance the classification performance.

Overall, the study demonstrates a promising approach to fine-tuning large language models for Bangla emotion analysis, achieving a reasonable accuracy given the dataset's challenges. However, there is ample room for improvement,

particularly in handling class imbalances and enhancing the precision and recall for minority classes.

5 Conclusion

In this study, we fine-tuned the `google/gemma-2b` language model to perform emotion analysis on a Bengali dataset. The dataset consisted of five emotion categories: Happy, Love, Sadness, Anger, and Fear. Through extensive data preprocessing, prompt engineering, and fine-tuning, we were able to achieve a reasonable level of performance with an overall accuracy of 71.17%.

Our results indicate that the model performs particularly well in detecting the 'Happy' emotion, achieving an F1-score of 0.81, but struggles with less represented classes such as 'Fear' and 'Anger', which is reflected in their significantly lower F1-scores. The imbalance in the dataset, where 'Happy' and 'Love' are dominant classes, likely contributed to this disparity in performance.

Despite these challenges, the model's ability to generalize to different emotions demonstrates the potential of using large language models for emotion detection tasks in low-resource languages like Bengali. Future work could focus on improving the model's performance on minority classes by incorporating techniques such as data augmentation, synthetic data generation, or more sophisticated balancing methods during training. Additionally, further fine-tuning with a larger and more balanced dataset could enhance the model's robustness and accuracy.

In summary, this work provides a foundation for emotion analysis in Bengali text, highlighting both the capabilities and limitations of current language models in handling nuanced emotional content in a low-resource language setting.

References

[Ras+24] Mohammad Rifat Ahmmad Rashid et al. "A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews". In: *Data in Brief* 53 (2024), p. 110052. ISSN: 2352-3409. DOI: `https://doi.org/10.1016/j.dib.2024.110052`. URL: `https://www.sciencedirect.com/science/article/pii/S235234092400026X`.