2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), 13-15 February 2025, CUET, Chattogram-4349, Bangladesh

2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) | 979-8-3503-5750-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ECCE64574.2025.11013999

Distinguishing Between Formal and Colloquial: A Multilingual BERT Approach to Bengali Language Classification

Sagor Aich[1], Koshik Debanath[2], Azmain Yakin Srizon[3]
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
Email: [1]sagor.aichh@gmail.com, [2]koshik.debanath@gmail.com, [3]azmainsrizon@gmail.com

***Abstract*—The Bengali language, rich in history and cultural significance, poses unique challenges in Natural Language Processing (NLP) due to its dual-register structure: Sadhu (formal) and Cholit (colloquial). These registers differ significantly in syntax, vocabulary, and usage, complicating tasks such as text classification, translation, and sentiment analysis. Language models not specifically trained to recognize these distinctions often misinterpret these variations, limiting the accuracy of Bengali NLP tools. To address this, a dataset from Mendeley was used to fine-tune the multilingual BERT (mBERT) model for distinguishing between Sadhu and Cholit registers. The fine-tuned model achieved an accuracy of 94.08%, effectively capturing the subtle lexical and syntactic differences between the two forms. This work advances Bengali NLP, enabling more precise applications in digital communication, automated translation, and linguistic analysis, while contributing to broader advancements in low-resource language processing.**

***Keywords*—Bengali Language Classification, Sadhu and Cholit Registers, Low-Resource Language Processing, Natural Language Processing (NLP), Transformer-Based Models, Multilingual BERT (mBERT).**

I. Introduction

The term 'Sadhu Bhasha' (Saint or Formal Bengali) refers to a more traditional and formal style, often used in classical literature and formal writing, and the term 'Cholit Bhasha' (Common or Colloquial Bengali) is a more conversational, modern form widely used in everyday speech and informal writing.

The Bengali language, one of the most widely spoken languages globally, exhibits two distinct registers: Sadhu (formal) and Cholit (colloquial). These linguistic variations carry unique grammatical structures, vocabulary, and stylistic elements, making the distinction between Sadhu and Cholit challenging for automated systems. Accurate classification of these registers is vital for improving the performance of natural language processing (NLP) applications in Bengali, including sentiment analysis, machine translation, and digital communication tools.

In recent years, neural network approaches, particularly those involving transformer-based models like the multilingual BERT (mBERT), have shown significant promise in language classification tasks across various languages. The BERT model's capacity to process contextualized embeddings of words allows it to capture the nuanced distinctions between Sadhu and Cholit forms in Bengali. This study seeks to leverage the mBERT model to classify Bengali sentences effectively into Sadhu or Cholit forms.

Prior studies have explored Bengali language classification through various methods. Traditional approaches have focused on rule-based and statistical models, often limited by their inability to handle complex language variations. Recent research has demonstrated that deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are more adept at managing such linguistic diversity. Moreover, specific studies have attempted to enhance Bengali text classification by employing neural networks trained on specialized Bengali corpora, achieving substantial improvements over conventional methods.

The availability of large, annotated Bengali datasets has been instrumental in advancing Bengali NLP. The dataset used in this study [1] provides a comprehensive collection of Bengali sentences categorized into Sadhu and Cholit, facilitating model training and evaluation. By fine-tuning mBERT on this dataset, we aim to assess its effectiveness in discerning between Sadhu and Cholit forms, contributing valuable insights to Bengali NLP.

This research applies mBERT to classify Bengali sentences into Sadhu and Cholit forms, leveraging recent advancements in transformer-based models. This approach holds significant implications for applications across Bengali language technology and contributes to the broader field of low-resource language processing.

II. Literature Review

This literature review examines key deep learning architectures relevant to Bengali text classification: Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Transformer models like BERT, and multilingual BERT (mBERT). These models have significantly advanced Natural Language Processing (NLP) and play a crucial role in addressing the complexities of Bengali language processing.

A. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are widely used for sequential data, such as text. They process input sequences element by element, maintaining a hidden state

that captures information from earlier time steps. This capability allows RNNs to model temporal dependencies in language, making them suitable for tasks like text classification, machine translation, and speech recognition [2]. However, standard RNNs are hindered by the vanishing gradient problem, which limits their ability to model long-range dependencies [3]. Despite this challenge, RNNs remain foundational in sequential data modeling and have been applied to various Bengali text classification tasks [4], [5].

$$h_t = \tanh(W_h x_t + U_h h_{t-1} + b_h)$$
$$y_t = \sigma(W_y h_t + b_y)$$

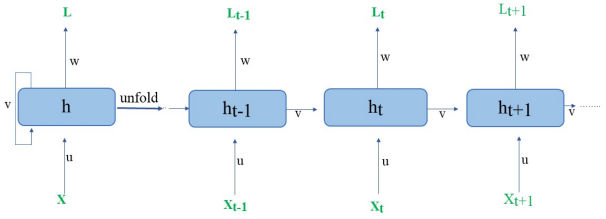

Figure 1. Architecture of Recurrent Neural Networks for Sequential Data Processing

B. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a specialized type of RNN designed to overcome the vanishing gradient problem through the use of memory cells [6]. These cells maintain a cell state that persists across long sequences, enabling LSTMs to model long-range dependencies more effectively than traditional RNNs. LSTMs are widely applied in NLP tasks such as machine translation and text classification [7]. In Bengali language processing, LSTMs have been utilized successfully for sentiment analysis [8] and parts-of-speech tagging [9]. However, their sequential nature leads to high computational costs when processing very long sequences.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot \tanh(c_t)$$

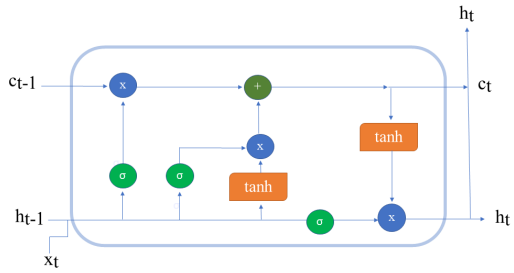

Figure 2. Long Short-Term Memory (LSTM) Network Architecture for Capturing Long-Range Dependencies

C. Transformer Models

Transformer models, introduced by Vaswani et al. [10], revolutionized NLP by replacing recurrent layers with self-attention mechanisms. The attention mechanism enables the model to evaluate the importance of words in a sequence regardless of their position, allowing parallel computation and significantly reducing training time compared to traditional RNNs and LSTMs. Transformers have consistently achieved state-of-the-art results across various NLP tasks and are increasingly applied in Bengali NLP, including named entity recognition and syntactic parsing [11]. Their architecture excels in capturing long-range dependencies and complex relationships, making them highly effective for classifying Bengali text.

D. BERT: Bidirectional Encoder Representations from Transformers

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model pre-trained on extensive text corpora, enabling it to capture rich contextual information [12]. Unlike earlier models, BERT is trained to understand bidirectional word context, resulting in superior performance on a wide range of NLP tasks. Through pre-training on large datasets and fine-tuning on task-specific data, BERT has set new standards for accuracy across various applications [13], [14]. In Bengali NLP, BERT has been applied to tasks such as technical domain classification [15], sentiment analysis [16], text classification, and question answering [17]. Its ability to capture contextual dependencies has made it a preferred choice for multilingual text classification tasks, including Bengali.

$$
\begin{aligned}
\mathbf{X}_0 &= \mathbf{E} + \mathbf{P}, \quad \mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \\
\text{MHA}(\mathbf{X}) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O, \\
\mathbf{F}(\mathbf{X}) &= \text{ReLU}(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \\
\mathbf{X}_{l+1} &= \text{LayerNorm}(\text{MHA}(\mathbf{X}_l) + \mathbf{X}_l), \\
\mathbf{X}_{l+1} &= \text{LayerNorm}(\mathbf{F}(\mathbf{X}_{l+1}) + \mathbf{X}_{l+1}), \\
\text{BERT}(\mathbf{X}) &= \mathbf{X}_L.
\end{aligned}
$$

In these equations, $\mathbf{X}_0$ represents the input embeddings, which combine token embeddings $\mathbf{E}$ and positional encodings $\mathbf{P}$. The self-attention mechanism, expressed as $\mathbf{A}$, calculates attention scores using query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices, with d_k denoting the dimension of the key/query vectors. The multi-head attention output (MHA) concatenates the results of multiple attention heads, followed by a linear transformation using the weight matrix $\mathbf{W}^O$. The feedforward layer applies non-linear transformations using learnable parameters $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$. Each transformer block employs layer normalization to stabilize training, with the final output $\mathbf{X}_L$ representing the contextualized embeddings generated by BERT.

E. Multilingual BERT (mBERT)

Multilingual BERT (mBERT) is a transformer-based model pre-trained on 104 languages, designed to handle multilingual NLP tasks without requiring individual models for each language. By capturing language-agnostic representations, mBERT is highly effective for cross-lingual transfer, where knowledge from one language aids tasks in another [18]. Applications of mBERT include cross-language text generation, hate speech detection, text classification [19], and sentiment analysis [20].

This capability is especially beneficial for under-resourced languages like Bengali, as mBERT enables effective text classification and language identification with minimal labeled data. Trained on large multilingual corpora, mBERT generalizes well across languages with shared syntactic or semantic features. Its bidirectional nature, inherited from BERT, allows it to capture rich contextual information from both the left and right contexts of a sentence.

In this study, we applied mBERT to a Bengali dataset to classify sentences as either Sadhu or Cholit, demonstrating its effectiveness in handling the linguistic diversity of Bengali.

III. Dataset Description

The dataset used in this study, titled "BanglaBlend: A Large-Scale Novel Dataset of Bangla Sentences Categorized by Saint (Sadhu) and Common (Cholit) Form of Bengali Language," contains a total of 7,350 instances, evenly distributed across two classes: Sadhu and Cholit. Each class comprises 3,675 instances, ensuring a balanced dataset. This balance is critical for effective model training, as it mitigates the risk of bias towards any specific class during the classification process.

Table II provides a summary of the data distribution between the two classes. Additionally, Figure 3 presents a word cloud visualization of the most frequent words in the dataset, highlighting key lexical patterns that distinguish the Sadhu and Cholit forms.

Figure 3. Visualization of Frequent Words in Dataset Sentences Using a Word Cloud

IV. Proposed Methodology

This section describes the methodology for classifying Bengali sentences into Sadhu and Cholit forms using a multilingual BERT (mBERT) model. The methodology encompasses data preparation, model initialization, training, and evaluation.

A. Data Loading and Splitting

The dataset is loaded from an Excel file containing Bengali sentences and their corresponding class labels (Sadhu or Cholit). Using the Pandas library, the data is read and prepared efficiently. To facilitate model training and evaluation, the dataset is split into training and testing subsets in an 80:20 ratio. Stratified sampling ensures balanced class representation across both subsets.

Let the total number of instances in the dataset be N, with N_{train} and N_{test} denoting the sizes of the training and testing subsets, respectively. The split can be mathematically expressed as:

$$N_{\text{train}} = 0.8N, \quad N_{\text{test}} = 0.2N$$

B. Model Initialization

The pretrained multilingual BERT model 'bert-base-multilingual-cased' is utilized for the classification task.

Table I
Sample Sentences Demonstrating Sadhu and Cholit Styles in Bengali

Category	Sample Sentences
Sadhu	'সেখানকার জানালা দিয়ে সমুদ্র দেখা যাইতেছিল', 'আমি কিছু দেখিতে পারিতেছি না', 'সকলেরই অনাবৃত দেহ সকলের সেই অনাবৃত বক্ষে আরশির ধুকধুকি চন্দ্রকিরণে এক একবার জ্বলিয়া উঠিতেছে', 'মেয়েটি সেদিন ভিক্ষুককে সাহায্য করিয়াছিল', 'তুমি প্রশংসা কর না কর বৃদ্ধ বসিয়া তোমায় পুরাতন কথা শুনাইবে'
Cholit	'নিজেকে আপনার মূল্যবোধের প্রতি সত্য থাকুন অন্য কারো জন্য আপনি কে আপস করবেন না', 'আমি গোসল করে খেতে যাব', 'সংবাদপত্রে স্বাধীনতা আন্দোলনের বিভিন্ন সংবাদ ছাপা হতো', 'তাকে বলা যায় টারজান জীবনের চেয়েও বড় একজন মহানায়ক', 'আমি এই ছুটিতে পরিবারের সাথে ঘুরতে গিয়েছিলাম'

Table II
Class Distribution of Dataset for Sadhu and Cholit Styles

Class	Number of Instances
Total Data	7350
Sadhu	3675
Cholit	3675

This model, trained on multiple languages, is highly suitable for handling Bengali text. The corresponding tokenizer is also loaded to preprocess the text data effectively. The model's architecture is initialized with parameters optimized for multilingual text classification tasks.

C. Data Tokenization

Tokenization converts sentences into numerical representations that the model can process. For a given input sentence S, the tokenizer maps S to a sequence of tokens:

$$\text{Tokens} = \text{Tokenizer}(S)$$

Each token is further mapped to an embedding vector. Sentences longer than the maximum token limit $L_{\max}$ are truncated, and shorter sentences are padded to ensure uniform input lengths. This ensures that all inputs have a consistent dimension L, where $L \leq L_{\max}$.

D. Dataset Creation

A custom dataset class is implemented to manage tokenized encodings $\mathbf{X}$ and their corresponding labels $\mathbf{y}$. This enables efficient access to individual samples during both training and evaluation phases. Each sample (x_i, y_i) represents a tokenized sentence and its associated label.

E. Training Configuration

The training process is configured with specific hyperparameters. The objective function $\mathcal{L}$ to minimize during training is the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c})$$

where: - N_{train} is the number of training samples, - C is the number of classes (Sadhu and Cholit), - $y_{i,c}$ is the ground-truth label for class c, - $\hat{y}_{i,c}$ is the predicted probability for class c.

Key hyperparameters include the number of epochs E, batch sizes B_{train} and B_{eval}, warmup steps W, and weight decay λ.

F. Model Training

The Hugging Face Transformers library's Trainer class is employed for model training. The optimizer, AdamW, is used to update model parameters, with the learning rate η dynamically adjusted using a linear scheduler. During training, the model learns to minimize the loss $\mathcal{L}$ and classify Bengali sentences accurately.

The update rule for a model parameter θ at step t is:

$$\theta_{t+1} = \theta_t - \eta_t \nabla \mathcal{L}(\theta_t)$$

where η_t is the learning rate at step t.

G. Evaluation

The trained model is evaluated on the test set to assess its performance. Metrics such as precision, recall, and F1 score are computed to quantify the model's ability to distinguish between Sadhu and Cholit forms accurately. Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. The metrics are calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This methodology integrates advanced neural network techniques, a balanced dataset, and a robust pretrained language model to effectively classify Bengali sentences. By leveraging these resources, the methodology achieves high accuracy in distinguishing between Sadhu and Cholit forms.

V. Experimental Analysis

This section presents the evaluation of the proposed approach for classifying Bengali sentences into Sadhu and Cholit forms. The model's performance was assessed using various metrics, including accuracy, precision, recall, F1 score, and loss curves. Additionally, a confusion matrix provides insights into the classification results.

The proposed model achieved an impressive accuracy of 94.08% on the test set, highlighting its effectiveness in distinguishing between the two forms of Bengali. These results demonstrate the robustness of the mBERT model in addressing complex text classification tasks in low-resource languages.

A. Model Performance

Table III summarizes the performance of different models for Bengali sentence classification. Metrics such as accuracy, precision, recall, and F1 score are included, with the mBERT model showing superior performance across all criteria. The mBERT model achieved an accuracy of 94.08%, significantly outperforming other models, including BiLSTM and Random Forest.

Table IV presents class-wise metrics for the mBERT model. The precision, recall, and F1 scores for both Sadhu and Cholit classes are high, with the model demonstrating balanced performance across both categories. The macro and weighted averages of these metrics confirm the overall effectiveness of the classification system.

B. Confusion Matrix

The confusion matrix, illustrated in Fig. 4, provides a detailed view of the classification results. It highlights the counts of true positives, true negatives, false positives, and false negatives for both classes. The matrix shows a balanced distribution, indicating the model's capability to accurately classify both Sadhu and Cholit forms.

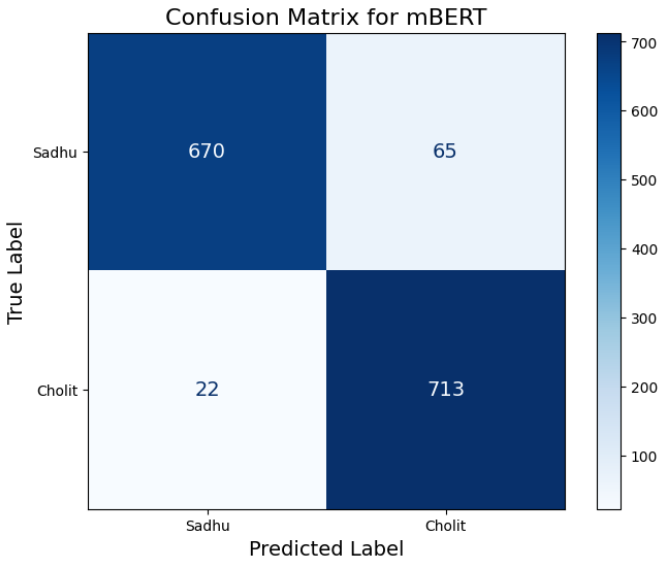

Figure 4. Confusion Matrix Depicting Classification Results for mBERT Model

C. Training and Validation Loss

Figures 5 and 6 display the training and validation loss curves, respectively. The training loss demonstrates a steady decline, indicating effective learning by the model during training. Similarly, the validation loss curve shows a consistent decrease, suggesting that the model is generalizing well to unseen data and avoiding overfitting.

D. Discussion

The experimental results demonstrate that the mBERT model significantly outperforms other models in accuracy, precision, recall, and F1 score. The high precision

Figure 5. Training Loss Progression Across Epochs

Figure 6. Validation Loss Progression Across Training Epochs

(98.65%) and F1 score (93.79%) highlight its ability to capture contextual nuances in Bengali text, making it particularly suited for this classification task. While models like BiLSTM and Random Forest show competitive performance, mBERT's superior metrics underscore its robustness and suitability for handling complex language variations.

Overall, the results validate the effectiveness of mBERT in classifying Bengali sentences with high accuracy and stability during both training and validation phases. These findings emphasize the potential of transformer-based models for advancing NLP applications in low-resource languages.

VI. Conclusion

This study introduced a methodology for classifying Bengali sentences into Sadhu and Cholit forms using a multilingual BERT model. The experimental evaluation achieved an impressive accuracy of 94.08%, highlighting the model's effectiveness in distinguishing between these two linguistic registers. The confusion matrix analysis indicates strong performance across both classes with minimal misclassifications. The steady decline observed in the training and validation loss curves confirms the model's ability to learn effectively from the training data while avoiding significant overfitting. These findings contribute significantly to the field of Bengali Natural Language Processing (NLP) by offering a robust framework for register

Table III
Performance Metrics of Various Models for Bengali Sentence Classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
K-Nearest	54.22	54.25	53.88	54.06
Cosine Similarity-Based	62.31	62.38	62.04	62.21
Centroid-Based Cos. Similarity	63.06	64.63	57.69	60.96
Decision Tree	72.24	69.17	80.27	74.31
LDA	74.69	76.89	70.61	73.62
XGBoost	79.12	81.66	75.10	78.24
Random Forest	80.68	81.28	79.73	80.49
SVM	81.43	81.91	80.68	81.29
Logistic Regression	81.97	84.97	77.69	81.17
BiLSTM	91.70	94.10	88.98	91.47
mBERT	**94.08**	**98.65**	**89.39**	**93.79**

Table IV
Detailed Class-Wise Performance Metrics for Bengali Sentence Classification

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Sadhu	96.82	91.16	93.90	735
Cholit	91.65	97.01	94.25	735
Macro avg	94.23	94.08	94.08	1470
Weighted avg	94.23	94.08	94.08	1470

classification. This framework has practical implications for applications such as automated translation, digital communication, and linguistic analysis. Expanding the dataset to include more diverse examples of Sadhu and Cholit forms could enhance model robustness. Further optimization could be achieved by exploring alternative architectures or fine-tuning methods. Applying this methodology to other South Asian languages with similar linguistic characteristics represents another promising direction for future research. The results demonstrate the feasibility and potential of using transformer-based models for tackling complex language classification tasks in Bengali, paving the way for advancements in multilingual NLP applications.

References

[1] U. Ayman, C. Saha, and Z. Mawa, "BanglaBlend: A Large-Scale Novel Dataset of Bangla Sentences Categorized by Saint (Sadhu) and Common (Cholito) Form of Bengali Language," 2024.

[2] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[3] Y. Bengio *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[4] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[5] Habib, M. Ahsan, and A. Akter, "Deep learning bangla text classification using recurrent neural network," vol. 8, pp. 10–16, 03 2022.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] S. Rahman and P. Chakraborty, "Bangla document classification using deep recurrent neural network with bilstm," in *Proceedings of International Conference on Machine Intelligence and Data Science Applications*. Singapore: Springer Singapore, 2021, pp. 507–519.

[8] A. Ahmed and M. Yousuf, *Sentiment Analysis on Bangla Text Using Long Short-Term Memory (LSTM) Recurrent Neural Network*, 01 2021, pp. 181–192.

[9] S. Sarkar and B. Anupam, "Part-of-speech tagging for bengali," 01 2007.

[10] A. Vaswani *et al.*, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[11] T. Alam, A. Khan, and F. Alam, "Bangla text classification using transformers," *arXiv preprint arXiv:2011.04446*, 2020.

[12] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[13] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, p. 1872–1897, Sep. 2020.

[14] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus : A survey of transformer-based pretrained models in natural language processing," 2021.

[15] K. Ghosh and A. Senapati, "Technical domain classification of bangla text using bert," 07 2021.

[16] A. Jana and A. Bhowmick, "Sentiment analysis for bengali using transformer based models," 12 2021.

[17] T. T. Mayeesha, A. M. Sarwar, and R. M. Rahman, "Deep learning based question answering system in bengali," *Journal of Information and Telecommunication*, vol. 5, no. 2, pp. 145–178, 2021.

[18] Y. Ma, "Cross-language text generation using mbert and xlm-r: English-chinese translation task." Association for Computing Machinery, 2024.

[19] A. Velankar, H. Patil, and R. Joshi, "Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi," in *Artificial Neural Networks in Pattern Recognition*. Cham: Springer International Publishing, 2023, pp. 121–128.

[20] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, "Multi-class sentiment analysis of urdu text using multilingual bert," *Scientific Reports*, vol. 12, 03 2022.