TransNeuNet: A Hybrid Transformer and Neural Network Model for Bangla Text Classification

Ronjon Kar, Md. Touhidur Rahman Limon, Tanjila Akter, Tausif Ahmed
Department of Computer Science & Engineering, East West University

Abstract

- Automated classification of Bengali news is necessary due to the rapid growth of online news and limited research in this area.
- A hybrid deep learning model, TransNeuNet, was developed, combining Bengali BERT and T5 embeddings with an ANN classifier.
- The dataset consists of 55,000+ articles from Kaggle and a custom web-scraped dataset, spanning 13 categories including Sports, Politics, Crime, Entertainment, and more.
- Preprocessing included stop-word removal, tokenization, label encoding, and date-time formatting to prepare the text for modeling.
- Experimental results show TransNeuNet achieved 93.53% accuracy, outperforming traditional models, and LIME was used to interpret predictions for transparency. Figure 1 illustrates the project's workflow.

Introduction/Background

- With the rapid growth of online news and social media, there is a need to automatically classify Bengali news to help readers and applications such as search engines, recommendation systems, and digital journalism analytics.
- Bengali is the 7th most spoken language in the world, with around 284 million speakers. Despite this, it receives limited attention in NLP due to a lack of high-quality datasets, complex word forms, sentence structures, and inadequate tools like lemmatisers and stemmers.
- Bengali news often contains mixed Bengali-English words, spelling variations, and informal writing, making tokenization, cleaning, and normalization difficult.
- A hybrid embedding system combining Bangla BERT and Bangla T5 was developed for robust feature representation. Multiple deep learning and transformer-based models were applied.
- A Streamlit web app in Figure 2 allows users to input Bengali articles and receive predictions quickly, making the system usable in real life.

Proposed Methodology

- Two Bengali news datasets—one from Kaggle and another created through web scraping—were combined, resulting in 13 final categories after filtering and merging. Figure 3 shows the class distribution.
- The data was cleaned by removing stopwords, non-Bangla text, numbers, URLs, and special characters, followed by tokenization and label encoding for model readiness.
- Class imbalance was corrected using under-sampling (5,000 samples per class), and exploratory analysis was performed through date-time conversion, heatmaps, and yearly/monthly article distributions.
- The final dataset of 55,333 articles was divided into 80% training and 20% testing, and text embeddings were generated using Bangla BERT (CLS representation) and Bangla T5 (mean-pooled encoder output).
- A hybrid model, TransNeuNet, was constructed by combining embeddings from BERT and T5, processing them through dense layers, and training the architecture (shown in Figure 4) with cross-entropy loss and the Adam optimizer on a Kaggle GPU environment.

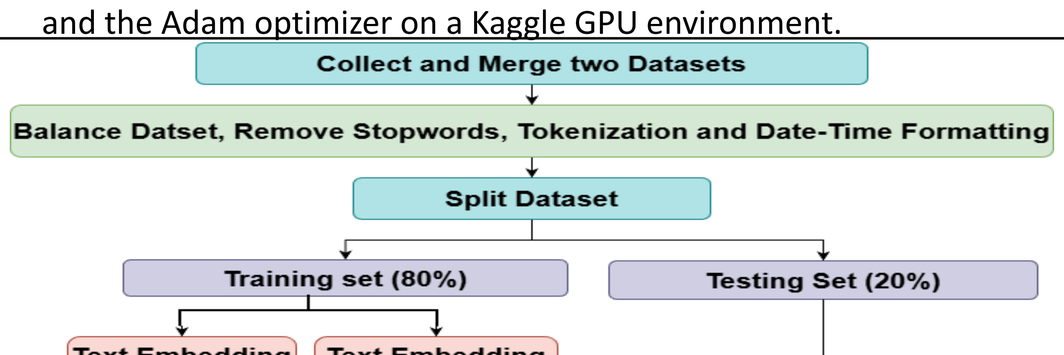

Figure 1: Methodology

Model	Accuracy	Precision	Recall	F1 score
CNN	0.7733	0.7781	0.7733	0.7740
BiLSTM	0.8835	0.8850	0.8835	0.8838
ANN	0.8889	0.8902	0.8889	0.8890
SVM	0.8747	0.8757	0.8747	0.8750
Decision Tree	0.6378	0.6383	0.6378	0.6379
Random Forest	0.8379	0.8429	0.8379	0.8391
TransNeuNet (Proposed)	0.9353	0.9359	0.9353	0.9354

Table 1: Performance Comparison of Different Models.

Experiments & Result Analysis

- The proposed TransNeuNet model achieved 93.53% accuracy with precision, recall, and F1-score around 0.93, outperforming all traditional ML and deep learning baselines.
- Models like Decision Tree (63.78%), Random Forest (83.79%), SVM (87.47%), CNN (77.33%), BiLSTM (88.35%), and ANN (88.89%) performed lower, showing that combining BERT + T5 embeddings gave TransNeuNet a clear advantage. Table 1 shows a comparison of the models.
- The confusion matrix in Figure 5 shows strong diagonal dominance, indicating correct predictions across 13 news categories. Most categories achieved an F1 score > 0.90, with sports, technology, and religion performing best.
- The loss curve shows stable convergence, and ROC curves show AUC values of 0.99–1.00, meaning the model can distinguish categories with almost perfect accuracy.
- LIME was used to explain predictions by highlighting important words that influenced classification. In Figure 6, Green bars show positive contributions, red bars show negative contributions, improving the transparency of the model's decisions.

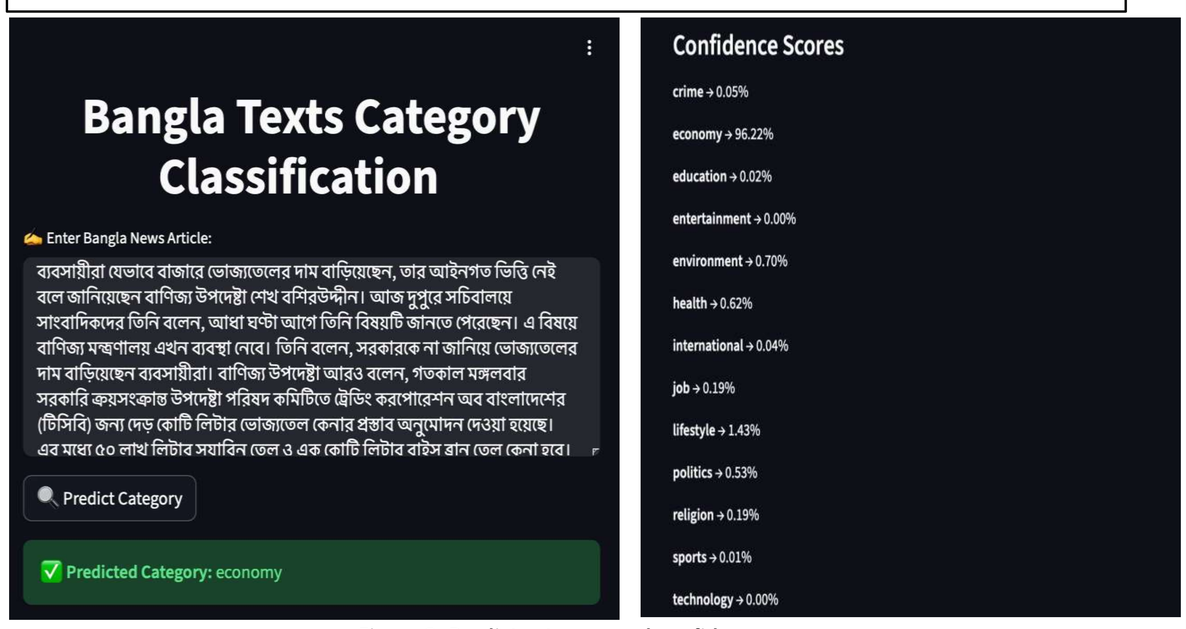

Figure 2. Predict category and confidence score.

Discussion

- All experiments were conducted on the Kaggle platform, which provided a free cloud environment with GPU support for efficient model training and testing.
- The model was trained using an NVIDIA Tesla P100 (16GB RAM, 13GB usable GPU memory) along with Python libraries such as PyTorch, TensorFlow, Scikit-learn, NumPy, and pandas.
- Since the model uses two transformer architectures (Bangla BERT and Bangla T5), it demands significant computational power and long training time.
- The model was trained on only 13 news categories, which does not fully represent real-world diversity; additionally, the target accuracy of 95%+ was not achieved.

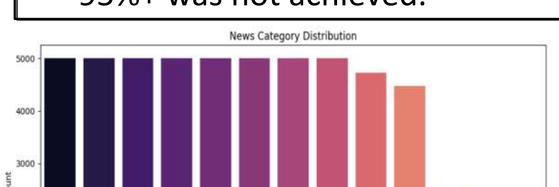

Figure 3. Balanced Class Distribution

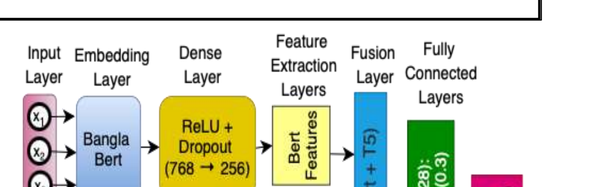

Figure 4. TransNeuNet Architecture

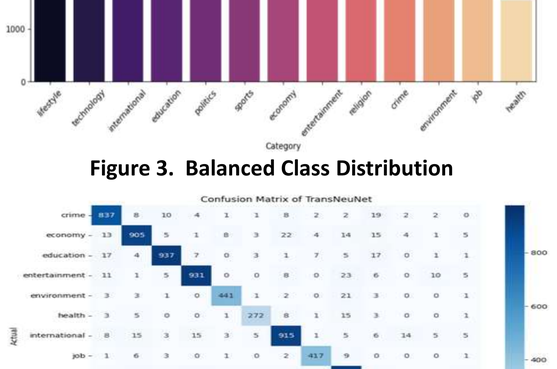

Figure 5. Confusion Matrix of TransNeuNet.

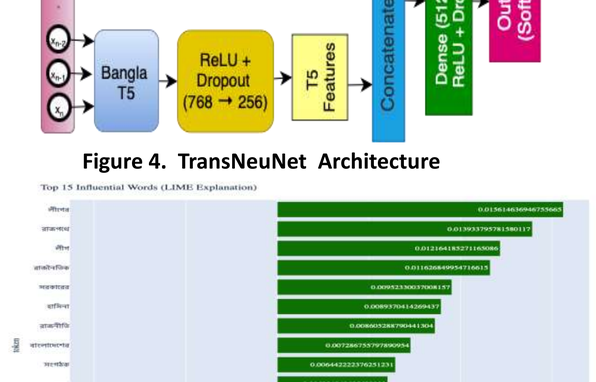

Figure 6. LIME Explanation

Conclusions

- The TransNeuNet model was developed, combining multiple machine learning and deep learning techniques for comparative analysis.
- Integration of Bangla BERT and Bangla T5 embeddings significantly improved classification performance.
- Interpretability was ensured using LIME, providing clear explanations for the model's category predictions.

References

[1] M. A. Z. K. Md. Mahbubur Rahman, "Bangla news classification using graph convolutional networks," Proceedings of the 2021 Inter national Conference on Computer Communication and Informatics (ICCCI), pp. 1–6, 2021.

[2] "Bengali language," Wikipedia, The Free Encyclopedia.

[3] M. I. H. Shakil Rana, "Newsnet: A comprehensive neural network hybrid model for efficient bangla news categorization," Proceedings of the 15th International Conference on Computing Communication Networking Technology (ICCCNT), pp. 522–527, 2024.

[4] P. T. Aysha Gazi Mouri, "An empirical study on bengali news headline categorization leveraging different machine learning tech niques," p. 312–317, 2022.

[5] S. S. Mohammad Rabib Hossain, "Different machine learning based approaches of baseline and deep learning models for bengali news categorization," International Journal of Computer Applications, vol. 176, pp. 10–16, 2020

EAST WEST UNIVERSITY — excellence in education — Department of Computer Science & Engineering