BOIGENRE: A Large-Scale Bangla Dataset for Genre Classification from Book Summaries

Rafi Hassan Chowdhury, Rahanuma Ryaan Ferdous
Islamic University of Technology
{rafihassan, rahanumaryaan}@iut-dhaka.edu

Abstract

The classification of literary genres plays a vital role in digital humanities and natural language processing (NLP), supporting tasks such as content organization, recommendation, and linguistic analysis. However, progress for the Bangla language remains limited due to the lack of large, structured datasets. To address this gap, we present BOIGENRE, the first large-scale dataset for Bangla book genre classification, built from publicly available summaries. The dataset contains 25,951 unique samples across 16 genres, showcasing diversity in narrative style, vocabulary, and linguistic expression. We provide statistical insights into text length, lexical richness, and cross-genre vocabulary overlap. To establish benchmarks, we evaluate traditional machine learning, neural, and transformer-based models. Results show that while unigram-based classifiers perform reasonably, transformer models, particularly BanglaBERT, achieve the highest F1-score of 69.62%. By releasing BOIGENRE and baseline results, we offer a valuable resource and foundation for future research in Bangla text classification and low-resource NLP. Our dataset and code is publicly available at https://github.com/Grumpy-Frog/BoiGenre

1 Introduction

Automatic classification of books by genre enhances digital library management and recommendation systems by aligning content with readers' interests. It enables personalized suggestions, improves user experience, and helps publishers understand audience preferences. However, most studies focus on English, leaving a significant gap for Bangla despite its global prominence. This gap stems from the lack of high-quality Bangla datasets, limiting the use of modern NLP techniques. Developing resources like BOIGENRE can bridge this gap and advance Bangla language research.

With the rise of modern machine learning frameworks, automated genre classification has become increasingly effective. This task involves identifying the genre of a book, movie, or artwork based on features such as the title, summary, or cover image. Previous studies have combined textual features like titles and summaries for genre prediction (Adhikari, 2025), while others explored language-specific datasets such as PPORTAL—the Portuguese literary dataset (Scofield et al., 2022). Moreover, genre detection has also been approached through visual data, as seen in the IMDb book cover dataset (Buczkowski et al., 2018).

Despite these developments, resources for Bangla literature remain limited. Existing datasets are often small or restricted to a few genres, reducing model generalization (Sethy et al., 2023). Furthermore, many studies have not leveraged transformer-based architectures such as BERT, which excel at contextual understanding. To address these gaps, we introduce the BOIGENRE dataset—a large-scale and diverse Bangla book genre dataset—and evaluate it using traditional machine learning, deep learning, and transformer models to establish strong baselines for future research.

In this paper, we introduce BOIGENRE, the first large-scale dataset for Bangla book genre classification based on summaries. This dataset addresses the scarcity of Bangla resources in natural language processing and reflects the diversity of Bangla literature. We evaluate multiple models, from traditional machine learning to transformer-based architectures, establishing strong baselines for future research. Future work may focus on expanding the dataset, incorporating full texts, and exploring advanced transformer or cross-lingual models to further enhance Bangla genre classification.

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 249–258

2 Literature Review

Most research on book genre classification has been conducted on English datasets such as Goodreads and Project Gutenberg. These studies primarily rely on user tags, full text, or book covers for genre detection (McAuley et al., 2017; Iwana et al., 2016). Some works have also used summaries, such as Bhuiyan et al. (Bhuiyan et al., 2023) and Adhikari (Adhikari, 2025), but these are based on small English datasets with limited genre coverage. In another direction, Pasha et al. (Pasha et al., 2023) focused on Bangla poem classification, though their work was restricted to only two genres and did not use book summaries. In addition, multilingual and cross-lingual approaches have been explored, including XLM-RoBERTa-based models for multiple languages (Classla, 2024), cross-lingual summarization (Zhang et al., 2024), and universal cross-lingual text classification (Savant, 2024), showing that genre detection can benefit from transfer learning across languages.

Over the years, methods for genre detection have evolved from simple TF-IDF and SVMs to deep learning models and transformer-based approaches like BERT. While these models show better performance, most prior works remain limited to English datasets and face issues such as label noise, class imbalance, and unclear genre definitions (Iwana et al., 2016; Kabir et al., 2023). In Bangla, resources are especially scarce, with existing datasets focusing mainly on sentiment analysis rather than genre classification (Kabir et al., 2023; Islam et al., 2021; Biswas, 2025; Alvi et al., 2022; Mahmud and Mahmud, 2024; Ahmed et al., 2023).

Our dataset, BOIGENRE, addresses these gaps by providing Bangla book summaries with well-defined genre labels. Unlike noisy user tags or cover images, summaries capture the narrative context more directly. This makes the dataset suitable for supervised genre detection and contributes a new resource for Bangla, which has been underrepresented in this research area (Kabir et al., 2023).

3 The BOIGENRE Dataset

In this section, we describe the process of curating the proposed BOIGENRE dataset. We outline the data sources, collection strategy, and refinement steps. These details provide the foundation for its use in downstream experiments.

3.1 Data Collection

We constructed our dataset by collecting information on books from the online bookstore Rokomari.[1] In total, we gathered metadata for 25,951 books spanning 16 distinct genres. The data collection process was carried out using the `BeautifulSoup` library. We first compile a list of URLs corresponding to the 16 genre categories available on the platform and then iteratively scraped the details of each book from these pages. Specifically, we scraped the *title*, *author*, *genre*, and *summary* fields, which together provide a comprehensive representation of the metadata of each book. This information forms the backbone of the dataset, as it allows for both high-level categorization across genres and fine-grained analysis based on author contributions and narrative content. By covering a wide range of genres and including summaries alongside bibliographic details, the dataset captures a diverse cross section of Bangla literature and creates opportunities for multiple downstream natural language processing tasks.

3.2 Preprocessing

Processing Step	Samples Remaining
Initial collection	98,811
After removing missing summaries	46,677
After removing duplicates	25,951

Table 1: Number of samples retained after each data pre-processing step.

The initial collection resulted in a total of 98,811 samples. However, a large portion of these records were incomplete, particularly missing summaries. After removing all samples without summaries, we retained 46,677 valid entries. Next, we identified and eliminated duplicate instances, where both the title of the book and the summary were identical across multiple records. This cleaning step was crucial to reduce redundancy and ensure dataset quality. Following the de-duplication process, the final dataset consisted of 25,951 unique book samples, which form the basis for our subsequent experiments and analyzes 1.

4 Statistical Analysis

Table 2 and related figures provide an overview of the statistical properties of the BOIGENRE dataset across genres. The distribution of samples reveals

[1] `https://www.rokomari.com`

Genre	N	Words / Summ.	Sents / Summ.	TTR	Vocab Size
Biography and Autobiography	5560	160.35±148.27	11.84±11.60	0.093	83148
Contemporary Novel	4669	136.17±89.32	14.21±10.61	0.103	65595
History and Tradition	3224	165.86±126.55	11.39±9.66	0.108	57626
Religious	2674	170.14±193.39	11.62±10.98	0.102	46290
Contemporary Story	2234	127.43±96.47	11.77±10.19	0.142	40506
Classic Novel	1189	139.59±112.08	13.56±11.91	0.183	30294
Thriller	1066	135.07±82.66	13.64±9.38	0.181	26125
Science Fiction	854	132.39±94.35	13.10±9.60	0.189	21348
Shishu Kishor	803	126.68±103.89	14.10±11.80	0.192	19567
Politics	778	172.64±159.49	11.36±10.34	0.175	23476
Philosophy	752	155.39±125.05	10.47±8.37	0.188	21991
Mystery	719	131.71±81.12	12.50±8.41	0.216	20497
Classic Story	593	135.25±189.75	12.08±21.28	0.246	19747
Adventure	354	140.39±132.46	14.01±12.03	0.261	12950
Math	267	178.46±171.03	11.67±10.93	0.178	8499
Cooking, Food and Nutrition	215	194.84±190.98	12.14±9.03	0.221	9274

Table 2: Summary statistics across genres. N is the number of summaries, *Words / Summ.* is the average number of words per summary (mean±std), *Sents / Summ.* is the average number of sentences per summary (mean±std), *TTR* is the type–token ratio (lexical diversity), and *Vocab Size* is the number of unique word types observed in each genre.

that *Biography and Autobiography* is the most well-represented category with 5,560 summaries, followed by *Contemporary Novel* with 4,669 and *History and Tradition* with 3,224. In contrast, specialized domains such as *Math* and *Cooking, Food and Nutrition* contain far fewer samples, only 267 and 215 respectively, indicating a strong imbalance across genres that realistically reflects the Bangla publishing landscape. Such imbalance also introduces challenges for downstream modeling, as classifiers must learn to generalize across both abundant and scarce categories without biasing predictions toward majority classes.

When looking at text length, notable differences emerge. Summaries in *Cooking, Food and Nutrition* and *Math* are the longest on average, exceeding 175 words per summary, while *Shishu Kishor* and *Contemporary Story* contain much shorter summaries, around 127 words. Interestingly, sentence-level statistics suggest that *Contemporary Novel* and *Shishu Kishor* have more segmented narrative structures, with around 14 sentences per summary, compared to only 10–12 sentences in most other genres. This aligns with the expectation that children's and contemporary fiction are written in shorter, more digestible units.

Lexical diversity, measured using type–token ratio (TTR), varies across genres. Genres with fewer samples, such as *Adventure* with a TTR of 0.261 and *Classic Story* with a TTR of 0.246, display the highest lexical diversity, indicating more varied vocabulary relative to their size. Larger domains like *Biography and Autobiography* with a TTR of 0.093 and *Religious* with a TTR of 0.102 show lower diversity, likely due to repeated thematic expressions. Broad domains such as *Biography and Autobiography* and *Contemporary Novel* contribute the largest vocabularies with over 65,000–83,000 unique types, while narrow domains like *Math* and *Cooking, Food and Nutrition* have smaller vocabularies, under 10,000 types.

These findings suggest that the BOIGENRE dataset captures a wide spectrum of Bangla literary styles, ranging from highly narrative forms to compact, domain-specific texts. At the same time, the inherent class imbalance poses a challenge, motivating the development of methods that can better handle skewed distributions and improve performance on minority genres.

5 Developing Benchmark for BOIGENRE

To provide a benchmark for Bangla literary text classification, we conduct experiments on the BOIGENRE dataset. The collection covers a wide spectrum of genres and exhibits substantial diversity in narrative style, linguistic expression, and vocabulary usage. Such variety makes BOIGENRE a suitable resource for testing the capacity of different NLP models to generalize across heterogeneous text domains.

Model	Feature Set	Vectorizer	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	Unigram	TF–IDF	63.58	62.88	63.58	63.02
Logistic Regression	Unigram	TF–IDF	61.00	63.10	61.00	61.40
Naive Bayes	Unigram	BoW	60.67	65.15	60.67	57.50
Gradient Boosting	Unigram	TF–IDF	55.47	55.04	55.47	53.54
SVM	Bigram	TF–IDF	54.10	52.88	54.10	53.21
SVM	2+3gram	TF–IDF	53.28	52.15	53.28	52.49
Naive Bayes	Bigram	BoW	55.84	57.15	55.84	52.05
Naive Bayes	2+3gram	BoW	54.76	55.54	54.76	51.11
Logistic Regression	2+3gram	TF–IDF	51.08	52.28	51.08	51.07
Random Forest	Unigram	TF–IDF	54.35	65.83	54.35	48.71
BiLSTM	Unigram	fastText	57.17	57.62	57.17	55.84
Custom CNN	Unigram	fastText	60.79	59.26	60.79	59.23
BanglaBERT	Contextual subword embeddings	BanglaBERT tokenizer	**69.34**	**70.26**	**69.34**	**69.62**

Table 3: Performance comparison of traditional machine learning models, deep learning architectures, and pretrained transformers for Bangla text classification. Accuracy, Precision, Recall, and F1 Score are reported as weighted averages across classes. Best results are highlighted in bold.

5.1 Baseline Models and Features

To establish benchmark results on the BOIGENRE dataset, we experiment with a diverse set of models, ranging from classical machine learning classifiers to neural architectures and pretrained transformers. This setup enables us to analyze the relative effectiveness of shallow lexical features versus contextualized embeddings in Bangla text classification.

For lexical features, we extract unigram and higher-order n-gram (bi- and tri-gram) representations at both the word and character levels. These features are vectorized using TF–IDF and Bag-of-Words (BoW), two widely used representations in text classification tasks. Classical classifiers, including Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Logistic Regression (LR) (Le Cessie and Van Houwelingen, 1992), Naive Bayes (NB) (John and Langley, 1995), Random Forest (RF) (Breiman, 2001), and Gradient Boosting (GB) (Friedman, 2001), are trained on these feature sets. These models serve as lightweight, interpretable baselines that capture surface-level lexical correlations.

To incorporate distributed representations, we utilize fastText embeddings (Joulin et al., 2017) as inputs to neural models. A BiLSTM (Hochreiter and Schmidhuber, 1997) is employed to capture long-range sequential dependencies, while a custom CNN (Kim, 2014) is used to detect local compositional patterns. Such architectures have shown effectiveness in morphologically rich languages, where subword-level modeling is essential.

We further evaluate a transformer-based model, BanglaBERT (Bhattacharjee et al., 2022), pretrained on large-scale Bangla corpora. As a contextualized encoder, BanglaBERT captures both semantic and syntactic information and represents the strongest baseline in our study.

For evaluation, we report Accuracy, Precision, Recall, and F1-score. Given the class imbalance in BOIGENRE, the weighted F1-score is emphasized as the primary metric, as it balances precision and recall across unevenly distributed categories (Sokolova et al., 2006).

5.2 Results & Findings

Table 3 reports the performance of traditional machine learning models, neural architectures, and pretrained transformers on the BOIGENRE dataset. *BanglaBERT* achieves the best results with 69.34% accuracy and 69.62% F1, outperforming both classical and deep learning baselines. Since the dataset is class-imbalanced, the F1 score is a more reliable indicator of performance, and BanglaBERT consistently leads across all metrics.

Among classical approaches, SVM with unigram and TF–IDF performs strongest with 63.58% accuracy, while higher-order n-grams degrade performance due to sparsity. Naive Bayes performs competitively on unigrams but shows weaker F1 scores, reflecting difficulty in handling minority genres. Tree-based ensembles such as Random Forest and Gradient Boosting perform worst, struggling with sparse, high-dimensional features.

Neural models using fastText embeddings, such as BiLSTM and a custom CNN, provide modest

Genre	Precision (%)	Recall (%)	F1 Score (%)
Adventure	40.85	50.00	44.96
Biography and Autobiography	79.52	77.05	78.26
Classic Novel	41.60	50.00	45.41
Classic Story	20.49	32.47	25.13
Contemporary Novel	75.19	67.05	70.89
Contemporary Story	60.53	66.59	63.41
Cooking, Food and Nutrition	100.00	92.86	96.30
History and Tradition	73.87	77.50	75.64
Math	96.49	96.49	96.49
Mystery	47.89	59.65	53.12
Philosophy	67.41	62.33	64.77
Politics	42.86	42.33	42.59
Religious	80.30	83.56	81.90
Science Fiction	73.17	83.33	77.92
Shishu Kishor	61.49	50.28	55.32
Thriller	63.01	52.87	57.50

Table 4: Classwise analysis for BanglaBERT model.

gains over traditional baselines, but remain behind BanglaBERT. Their limitations suggest that sequential or convolutional architectures alone cannot capture the full morphological and semantic complexity of Bangla text.

The superiority of BanglaBERT can be attributed to its pretraining on large-scale Bangla corpora, which enables it to model polysemy, discourse-level structure, and genre-specific variation. At the same time, the relatively strong unigram-based linear classifiers highlight that surface lexical features remain effective for Bangla classification tasks. Together, these findings emphasize both the promise of pretrained transformers and the continued value of lightweight baselines for future work on Bangla NLP.

5.3 Error Analysis

As shown in Table 4, BanglaBERT achieves strong performance on high-resource genres such as Biography and Autobiography (F1 78.26) and Religious (F1 81.90), as well as on domain-specific but lexically narrow categories like Math and Cooking, Food and Nutrition, where F1-scores exceed 95. In contrast, minority genres with high lexical diversity such as Adventure with F1 44.96, Classic Story with F1 25.13, and Mystery with F1 53.12 remain more difficult. This highlights how class imbalance and high type–token ratio jointly limit generalization. Table 2 shows that Adventure has a TTR of 0.261 and Classic Story has a TTR of 0.246, indicating strong lexical variation despite very small sample sizes, which makes them harder to model consistently.

The word-level statistics in Appendix Table 6 reveal that frequent unigrams such as করে, এই, তার, and এক appear across many genres, offering limited discriminative value. Their dominance leads to frequent misclassification between closely related categories, most notably Contemporary Novel, Classic Novel, Classic Story, and Shishu Kishor. Appendix Figure 1 confirms this pattern, showing cosine similarity values above 0.90 among these genres, which indicates a high degree of lexical overlap. Such overlap makes surface word distributions insufficiently distinctive and explains why classifiers often confuse them. In contrast, domain-specific genres like Math and Cooking, Food and Nutrition exhibit much lower similarity with other categories, with cosine scores below 0.55. Despite having fewer samples, these genres achieve higher classification accuracy because their specialized vocabularies are more distinctive. These findings show that classification errors in BOIGENRE are shaped not only by class imbalance but also by substantial cross-genre vocabulary overlap, underscoring the need for models that can capture deeper semantic and contextual information beyond surface lexical similarity.

Genre	Sample Count
Biography and Autobiography	294
Politics	123
History and Tradition	41
Contemporary Story	3
Philosophy	2

Table 5: Genre-wise distribution of books associated with "বঙ্গবন্ধু" in the BOIGENRE dataset.

Another source of difficulty comes from thematic overlap across genres. Table 5 shows that

books related to বঙ্গবন্ধু are distributed across Biography and Autobiography, Politics, and History and Tradition. Because thematically similar material is spread across multiple genres, the boundaries between classes are blurred, making it harder for the model to assign the correct label.

Taken together, these observations show that misclassification in BOIGENRE is shaped not only by class imbalance but also by high lexical diversity in low-resource genres and significant vocabulary overlap across thematically related categories. Future work can explore class-aware training strategies, targeted data augmentation, and representation learning approaches that capture deeper semantic and discourse-level patterns beyond surface lexical similarity.

6 Conclusion

In this paper, we introduce the first large-scale dataset BOIGENRE for classifying the genres of Bangla books based on summaries. This dataset represents an important step toward enriching Bangla resources in natural language processing and capturing the diversity of Bangla literature. We further evaluate a range of traditional, deep learning, and transformer-based models to establish strong baselines for future research. Future work may focus on expanding the dataset with additional sources and genres, incorporating full texts for deeper contextual understanding, and addressing class imbalance and linguistic variations. Moreover, exploring human evaluation methods and advanced transformer or cross-lingual models could further enhance the robustness and generalization of Bangla genre classification systems.

7 Limitations

Despite the contributions of this study, several limitations remain. The BOIGENRE dataset was compiled solely from the Rokomari website, which may not capture the full diversity of Bangla literature in terms of writing style, era or publication source. The dataset also exhibits class imbalance, as certain genres contain fewer samples than others, potentially affecting the model's ability to generalize across all categories. Furthermore, since the dataset is based only on book summaries rather than complete texts, it may lack deeper contextual or stylistic cues that are often essential for accurate genre identification. The current work focuses on sixteen primary genres, leaving out finer subgenres or overlapping categories that could provide more nuanced classification. Linguistic variations, such as regional dialects, informal spellings, and transliterated words, can also introduce inconsistencies within the data. In addition, the evaluation relied solely on quantitative performance metrics like Accuracy, Precision, Recall, and F1-score, without incorporating human judgment or interpretability analysis. Lastly, while transformer-based models such as BanglaBERT achieved strong performance, further experiments with multilingual or domain-specific transformers could provide deeper insights into model robustness and cross-domain generalization.

References

Jhimli Adhikari. 2025. Leveraging book genre classification using machine learning. *DESIDOC Journal of Library & Information Technology*, 45(3).

Zishan Ahmed, Shakib Sadat Shanto, and Akinul Islam Jony. 2023. Advancement in bangla sentiment analysis: A comparative study of transformer-based and transfer learning models for e-commerce sentiment classification. *Journal of Information Science and Engineering*, 15(6):48–63.

Nasif Alvi, Kamrul Hasan Talukder, and Abdul Hasib Uddin. 2022. Sentiment analysis of bangla text using gated recurrent neural network. In *International Conference on Innovative Computing and Communications*, pages 77–86. Springer.

Abhik Bhattacharjee, Shammur Absar Sarker, and et al. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7767–7778.

S.R. Bhuiyan, M.R.H. Khan, U.S. Afroz, and M.S. Rahman. 2023. Identifying genre of a book from its summary using machine learning approach. In *Proceedings of the International Conference on Advances in Computing, Communication and Control*. Springer.

Jahanur Biswas. 2025. Bangla sentiment dataset. Mendeley Data.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Przemyslaw Buczkowski, Antoni Sobkowicz, and Marek Kozlowski. 2018. Deep learning approaches towards book covers classification. In *ICPRAM*, pages 309–316.

Classla. 2024. Xlm-roberta base multilingual text genre classifier. Accessed: 2025-10-01.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Khondoker Islam and 1 others. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. Kaggle.

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*.

George H John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 427–431.

M. Kabir and 1 others. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. https://arxiv.org/abs/2305.12053.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Saskia Le Cessie and Hans C Van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, pages 191–201.

Hemal Mahmud and Hasan Mahmud. 2024. Enhancing sentiment analysis in bengali texts: A hybrid approach using lexicon-based algorithm and pre-trained language model bangla-bert. *arXiv preprint arXiv:2411.19584*.

Julian McAuley and 1 others. 2017. Goodreads book graph datasets. https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html.

Syed Tangim Pasha, Ashraful Islam, Mohammed Masudur Rahman, Eshtiak Ahmed, and Zahangir Alam. 2023. Genre classification of bangla poem using machine learning and deep learning techniques. Technical report, Independent University, Bangladesh.

R. Savant. 2024. Universal cross-lingual text classification. *arXiv preprint arXiv:2406.11028*. Accessed: 2025-10-01.

Clarisse Scofield, Mariana O Silva, Luiza de Melo-Gomes, and Mirella M Moro. 2022. Book genre classification based on reviews of portuguese-language literature. In *International Conference on Computational Processing of the Portuguese Language*, pages 188–197. Springer.

Abhisek Sethy, Ajit Kumar Rout, Archana Uriti, and Surya Prakash Yalla. 2023. A comprehensive machine learning framework for automated book genre classifier. *Revue d'Intelligence Artificielle*, 37(3):745.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021.

R. Zhang and 1 others. 2024. Cross-lingual cross-temporal summarization: Dataset creation, modeling, and evaluation. *Computational Linguistics*, 50(3):1001–1023.

A Appendix

Genre	10 Most Frequent Words
Adventure	এক (491), করে (404), থেকে (361), আর (334), তার (311), এই (260), কিন্তু (200), একটা (197), এবং (192), করতে (172)
Biography and Autobiography	এই (7773), এবং (6355), তিনি (6264), করে (6255), তাঁর (5892), থেকে (5171), তার (4503), এক (3964), এর (3804), আর (3329)
Classic Novel	করে (1572), তার (1371), এই (1350), এক (1082), আর (1056), থেকে (895), আমার (774), এবং (716), কিন্তু (693), আমি (619)
Classic Story	গল্প (873), এই (693), করে (638), তার (540), এক (484), আর (453), থেকে (372), এবং (339), গল্পের (324), আছে (283)
Contemporary Novel	করে (6746), তার (5985), এই (5265), এক (4520), আর (4509), থেকে (3735), হয় (2771), হয়ে (2681), আমার (2498), কিন্তু (2489)
Contemporary Story	গল্প (3123), করে (2621), এই (2367), তার (1736), আর (1638), এক (1547), থেকে (1479), গল্পের (1456), আমার (1410), জীবনের (1259)
Cooking, Food and Nutrition	রান্না (438), এই (357), রান্নারের (329), করে (303), খাবার (229), চিকেন (225), করা (222), থেকে (204), ইলিশ (199), এবং (195)
History and Tradition	এই (4809), এবং (4253), করে (3973), ইতিহাস (3396), থেকে (3340), একটি (2227), এর (2119), ছিল (2078), করা (2077), হয়েছে (1964)
Math	গণিত (831), গণিতের (504), এই (501), করে (398), এবং (351), করা (333), বইটি (321), জন্য (299), থেকে (272), সমাধান (271)
Mystery	এক (851), করে (842), এই (821), আর (654), তার (650), থেকে (572), রহস্য (499), কিন্তু (459), সেই (414), আছে (351)
Philosophy	এবং (1150), এই (1116), করে (929), থেকে (649), করা (583), দর্শন (538), তার (536), একটি (497), হয়েছে (473), এর (454)
Politics	এই (1157), এবং (1110), করে (1067), থেকে (734), একটি (634), তিনি (590), করা (583), রাজনৈতিক (561), তার (548), শেখ (535)
Religious	করে (3904), এবং (3412), করা (2988), থেকে (2901), এই (2887), জন্য (2424), এর (2349), তার (2213), আমাদের (1892), আর (1701)
Science Fiction	করে (1208), তার (971), এই (919), এক (898), আর (854), থেকে (817), করতে (528), কিন্তু (481), একটা (455), এবং (443)
Shishu Kishor	করে (1154), তার (908), আর (834), এই (727), এক (597), থেকে (592), কিন্তু (448), একটা (439), কথা (433), আমার (408)
Thriller	এক (1708), করে (1323), তার (1316), এই (1255), আর (1018), থেকে (916), কিন্তু (747), করতে (641), একটা (579), হয়ে (579)

Table 6: Top 10 most frequent Bangla word unigrams for each genre in the BOIGENRE dataset, shown with their raw occurrence counts. The list highlights the strong lexical overlap across narrative genres such as *Contemporary Novel*, *Classic Novel*, and *Shishu Kishor*, where frequent tokens like করে, এই, তার, and এক appear repeatedly across categories. This overlap contributes to cross-genre confusion observed in later analyses (Table 4 and Figure 1).

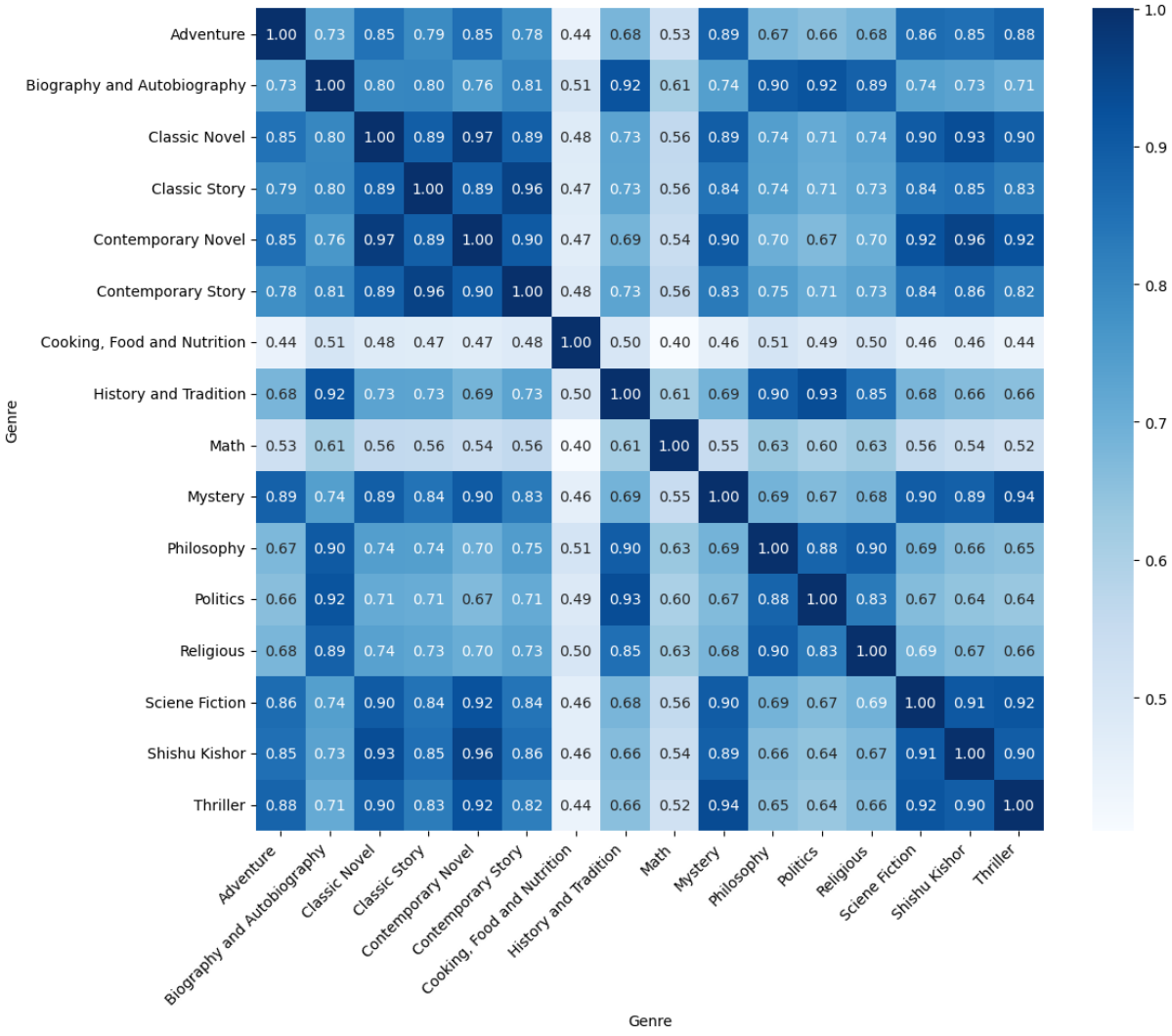

Figure 1: Cosine Similarity of Vocabulary Between Genres (TF-IDF Vocabulary).

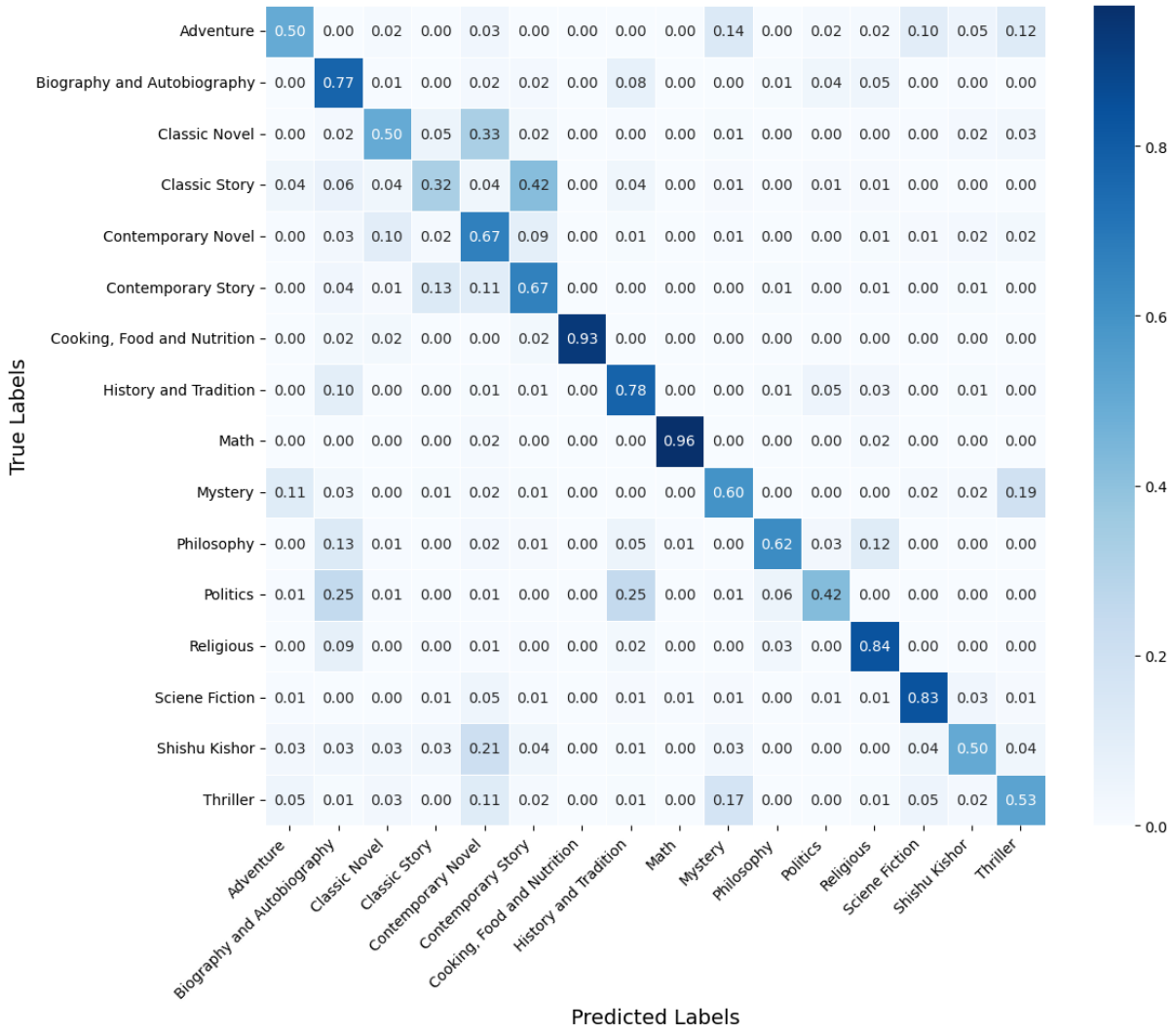

Figure 2: Normalized confusion matrix for BanglaBERT on the BOIGENRE dataset. The diagonal values indicate correctly predicted proportions per genre, while off-diagonal values represent misclassifications between closely related categories.