BWSD: A Bangla Web-based Summarization Dataset

Anonymous ACL submission

Abstract

In the last decade, natural language processing (NLP) has gained significant interest as it helps simplify human tasks and fulfills the desire to communicate with computers using human (natural) language. Yet, Bangla abstractive text summarization remains underexplored despite its widespread use. This work introduces Bangla Web-based Summarization Dataset (*BWSD*), a publicly available web-sourced Bangla summarization dataset, comprising 1,100 documents, alongside a custom preprocessing module for Bangla text processing. We propose a Bangla abstractive text summarization system, a freely available Bangla abstractive text summarization system, to evaluate BanglaT5, BanglaBERT, and mT5, with mT5 achieving the highest ROUGE-2 score (22.57). Despite challenges in data availability and linguistic complexity, our approach generates coherent, concise summaries, providing essential resources for Bangla NLP research. The dataset[1], the custom preprocessing module[2], and the system[3] are publicly available.

1 Introduction

Text summarization is a core task in natural language processing (NLP), aiming to shrink large volumes of text into concise, informative summaries while maintaining fundamental content. With the exponential growth of digital text, automatic summarization has become necessary in applications such as search engine optimization, news collection, and academic research. High-resource languages like English have seen notable advancements, benefiting from large-scale datasets and refined Transformer-based architectures that deliver high-quality summaries. These models effectively handle both extractive and abstractive summarization, producing coherent and contextually correct summaries across diverse domains.

However, Bangla, despite being the seventh most spoken language globally with over 273 million speakers, remains under-represented in summarization research due to the lack of annotated datasets, pre-trained models, and standardized processing tools (Goswami et al., 2023). Existing Bangla summarization efforts are largely domain-specific and rely on rule-based or extractive techniques, limiting their adaptability to diverse content. Addressing these challenges, this paper presents the Bangla Web-based Summarization Dataset (*BWSD*) - a publicly available, web-sourced dataset containing 1,100 diverse documents. Also, we introduce a custom preprocessing tool to enhance text preprocessing. Our experiments fine-tune and evaluate several pre-trained large language models (LLMs), including BanglaT5 (Bhattacharjee et al., 2022), BanglaBERT (Bhattacharjee et al., 2021a), and Multilingual T5 (mT5) (Xue, 2020), for Bangla abstractive summarization. Among them, mT5 achieves the highest ROUGE-2 score of 22.57, showing it as a strong candidate for Bangla summarization. This research contributes to addressing the resource gap in Bangla NLP by providing essential datasets, tools, and benchmark results to facilitate future advancements in the field.

2 Existing Datasets

Several datasets exist for Bangla text-to-text summarization, including XL_Sum (Hasan et al., 2021), Bengali abstractive news summarization (BANS) (Bhattacharjee et al., 2021b), BanglaSum (Roy, 2024), LR_Sum (Palen-Michel and Lignos, 2022), Bengali Text Summarization dataset (BTS), and Bangla Natural Language Processing Corpus (BNLPC). Some existing studies have utilized these data sets, i.e. XL_Sum ((Hasan et al., 2021; Shahariar et al., 2023; Ahmed et al., 2024)), BANS

[1] https://doi.org/10.5281/zenodo.14702674
[2] https://github.com/BWSDataset/readiness
[3] bwsdataset.github.io

((Sultana et al., 2022; Mukherjee, 2022; Miazee et al., 2025)), BTS ((Singha and Rajalakshmi, 2023)), and BNLPC ((Hasan et al., 2023; Rony and Islam, 2024; Khan et al., 2023)). XL_Sum is a multilingual dataset with 10,126 Bangla articles, a total word count of 1.2 million, and a unique word count of 155,276. Article lengths range from 400 to 1,800 words, while summaries are between 20 and 120 words. BANS comprises 19,096 Bangla articles with a total word count of 857,341 and 44,318 unique words. Articles range from 5 to 76 words, and summaries are 3 to 12 words. BanglaSum includes 9,311 Bangla articles, totaling 6.5 million words, with 185,376 unique words. Its articles span 14 to 3,500 words, and summaries range from 1 to 1,428 words. LR_Sum, a smaller multilingual dataset, features 715 Bangla articles with a total word count of 27,288, 10,583 unique words, article lengths of 150 to 450 words, and summaries of 25 to 45 words. Appendix A presents the comparative analysis of *BWSD* with the existing state-of-the-art dataset.

However, these datasets are content-specific, focusing only on news articles. To support broader applications, a general web-based Bangla dataset for text summarization is needed.

3 Bangla Web-based Summarization Dataset

This work addresses the aforementioned problems by presenting a large corpus of Bangla documents. One of the major contributions of this paper is the creation of a Bangla article dataset, *BWSD*: Bangla Web-Based Summarization Dataset. *BWSD* consists of 1,100 articles containing more than 28,000 words along with their corresponding summaries. This publicly available dataset can serve as a ground-truth resource for various Bangla natural language processing (BNLP) applications, such as summarizing social media data, news articles, and blog posts. The proposed dataset encompasses diverse text categories with a rich and varied vocabulary. This dataset is expected to facilitate research in BNLP significantly.

The subsequent subsections provide a detailed overview of the creation process of the proposed *BWSD* dataset, present a quantitative analysis of *BWSD*, conduct a qualitative study of the datasets used in our experiments, and conclude with a discussion of the key findings of our experiments.

3.1 Creation Process

Our *BWSD* creation procedure follows a carefully designed process to ensure reliability and usefulness. It starts with manually collecting articles from various sources, including online newspapers, blogs, and social media (Facebook). These source variations help us capture various writing styles and content types, ensuring that the dataset reflects a broad spectrum of Bangla texts. The collected content is compiled into an Excel file, creating the initial foundation for further processing. Figure 1 presents a workflow of our proposed *BWSD* dataset creation process.

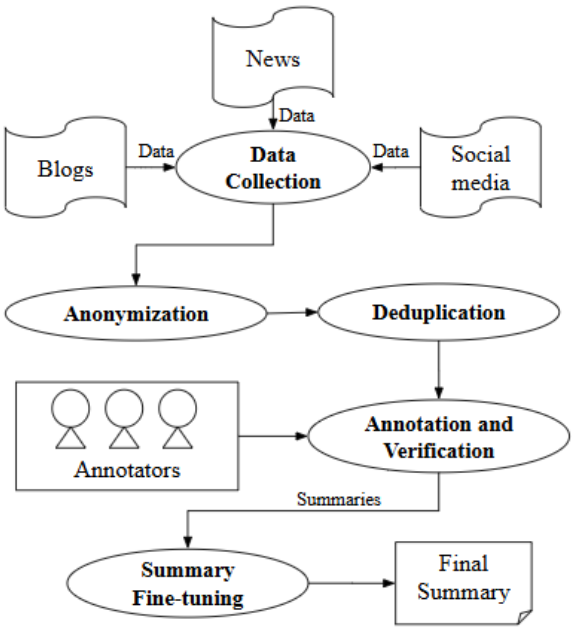

Figure 1: *BWSD* dataset creation process

Figure 1 shows once the articles are collected, the data undergoes two important preprocessing steps. The first is *Anonymization*, where sensitive or personal information is removed to maintain privacy. The second step *Deduplication* filters duplicate entries to ensure cleaning and avoiding redundancy of the dataset. These preprocessing efforts are essential for maintaining the quality of *BWSD*, making it well-prepared for research and practical applications.

After preprocessing, the summarization process begins. Three native Bangla speakers independently read and summarized each article, creating multiple summaries based on their understanding. This approach allows for a more natural and human-like representation of the text. By relying on human annotators, *BWSD* ensures that the summaries are

not only accurate but also contextually meaningful. This summarization process is widely used in linguistic research and helps maintain consistency with established practices in language analysis.

To further enhance the quality of *BWSD*, the annotation process includes inter-annotator agreement (IAA) measures. Each annotator works independently, and their summaries are compared using a Cosine similarity matrix. This similarity matrix evaluates how closely the summaries align, ensuring consistency across annotations. If discrepancies are found, the summaries are refined through an iterative process. For example, the similarity between the first two independent summaries is calculated, leading to the generation of a fourth summary. This fourth summary is then compared with the third independent summary to produce a final summary. The final product of this process, referred to as the final summary, represents the most concise and accurate summary version of each article.

The rigorous process used to create *BWSD* ensures that it is a reliable resource for BNLP research. Its summaries are carefully crafted and refined, making them suitable for evaluating BATSumm models. The combination of human annotation and computational techniques, like the Cosine similarity matrix, gives the dataset a strong foundation for supporting linguistic and NLP studies.

3.2 Quantitative Analysis

This section presents the quantitative analysis of *BWSD*. It provides an overview of *BWSD*, highlighting its size, domain diversity, and key characteristics. It consists of 1,100 articles with a total word count of 283,573 and 23,979 unique words. The articles range in length from 100 to 1,400 words, while the summaries are between 25 and 135 words. All of the summaries are crafted using an abstractive summarization approach. The dataset spans three categories, i.e. news, blogs, and social media posts. The news category includes 200 articles each from Prothom Alo, Kaler Kontho, and Samakal and 100 articles from Bangla News 24 newspaper. In the blog category, 200 articles are collected from Muktomona and 100 articles from Tarunyo. Additionally, 100 social media posts are sourced from Facebook. The dataset is entirely in Bengali and offers a diverse range of content, making it suitable for developing and evaluating Bangla abstractive text summarization (BATSumm) systems across various text types.

3.3 Qualitative Analysis

To evaluate the quality of proposed *BWSD*, We proposed a Bangla abstractive text summarization system (BATSS) employing LLMs to summarize Bangla text. Subsequent sections present the quality evaluation process of the proposed system.

3.3.1 BATSS Framework

BATSS uses several pre-trained LLMs to assess their effectiveness in the BATSumm domain. Specifically, this system employs BanglaT5, BanglaBERT, and MT5 experimenting with their advanced language understanding capabilities. Figure 2 presents the overview of the proposed BATSS methodology.

To prepare the data for effective text summarization, we perform several preprocessing steps, including *Normalization*, Custom data preprocessing, and *Tokenization*. We employ a standard normalizer (Hasan et al., 2020) for *Normalization*, our proposed data preprocessing module for further preprocessing, and LLM-specific tokenizer for *Tokenization*. Hyper-parameter tuning allows us to fine-tune such parameters as learning rate, training batch size, evaluation batch size, weight decay, and number of epochs altogether for optimal performance of this model. The hyper-parameter settings for the proposed BATSumm system are shown in Appendix B.

3.3.2 Ablation Study

We test each stage of our proposed architecture to demonstrate the usefulness of including those steps in our model. Table 1 presents the ablation study of our proposed dataset and proposed architecture on different data preprocessing steps. The study evaluates the performance of these methods using the ROUGE metric, which is a standard measure for assessing the quality of summaries. The ROUGE scores are reported for three variants: R-1, R-2, and R-L, which measure the overlap of unigrams, bigrams, and the longest common subsequence between the generated summary and the reference summary, respectively.

From Table 1, we observe some key notes of our rigorous experiments. Those are: 1) Normalization generally improves the ROUGE scores across all methods, indicating that standardizing the text enhances summarization quality. 2) Our proposed custom preprocessing further boosts performance, suggesting that preparing the text with a data preprocessing tool for summarization is beneficial. Fi-

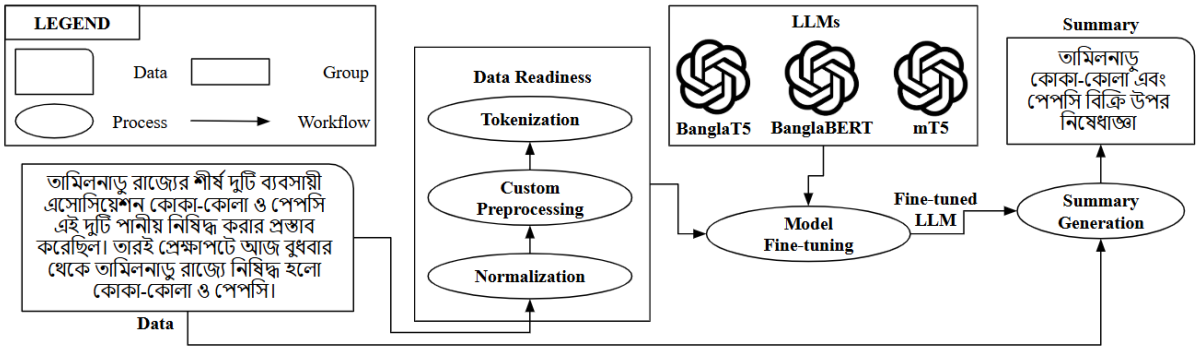

Figure 2: Proposed BATSS methodology

nally 3) The combination of Normalization and the proposed custom preprocessing module consistently yields the best results, demonstrating the effectiveness of using both preprocessing techniques together.

Method	ROUGE Score		
	R-1	R-2	R-L
BanglaT5			
C1	18.86	9.74	15.37
C2	22.75	11.26	19.46
C3	25.28	12.83	20.19
C4	25.36	**12.97**	21.09
BanglaBERT			
C1	24.72	9.08	21.18
C2	26.94	11.64	23.73
C3	36.13	14.37	27.21
C4	38.01	**16.95**	31.76
Multilingual T5			
C1	33.62	17.75	28.16
C2	34.59	18.53	28.94
C3	38.47	22.03	31.86
C4	41.68	**22.57**	32.28

Table 1: Ablation study of different data preprocessing techniques (C1: *BWSD*, C2: *BWSD* + Normalization, C3: *BWSD* + Custom preprocessing module, C4:*BWSD* + Normalization + Custom preprocessing module)

3.4 Discussion

We also evaluated the quality of existing state-of-the-art datasets using our system and within our experimental settings, ensuring a consistent and fair comparison across different datasets and methodologies. Each model's performance is measured using rouge matrices, which assess summarization quality based on n-gram overlap. The results show that *BWSD* consistently outperforms previously utilized datasets, achieving the highest scores across all three models. We observe that mT5 achieves the best accuracy among those LLMs. It achieves R-1, R-2, and R-L scores of 41.68, 22.57, and 32.28, respectively. It is not the case that all model always produces the best results using our proposed dataset. In the case of the BanglaBERT model, it produces better results with the BNLPC dataset, particularly in the R-2 score of 17.21. It also shows decent results, 16.95 R-2 score with *BWSD*. On the other hand, BanglaT5 shows low rouge scores but still benefits significantly from the proposed dataset. Appendix C shows that our proposed dataset provides better summaries across all models. High R-2 scores indicate better coherence, making summaries more informative. These findings highlight the importance of quality datasets and confirm mT5 as the most effective model.

4 Conclusion

This paper presents the Bangla web-based summarization dataset, (*BWSD*) containing 1,100 texts and a custom preprocessing module to enhance Bangla text preprocessing for abstractive summarization. Our proposed system, BATSS leverages several pre-trained LLMs, among those the mT5 achieved the best performance with a ROUGE-2 score of 22.57. Despite the complexities of Bangla grammar and the limitations of data availability, our work demonstrates significant progress. Also, BATSS generates concise and coherent summaries from existing datasets. Future efforts will address existing research gaps, focusing on dataset expansion, variable-length summarization, and advanced modeling techniques. This research marks a significant contribution to Bangla NLP by providing a publicly available dataset, a custom preprocessing tool, and a comprehensive evaluation of LLMs for Bangla summarization.

Limitations

Though our work introduces advancement in the BATSumm, several limitations remain in our work. First, while *BWSD* is one of the web-annotated publicly available Bangla summarization dataset, it still lacks the scale of high-resource language datasets, which limits our model generalization. Second, the number of articles in *BWSD* is relatively small. This limitation hinders the performance improvement of our model. Third, the quality of our summaries is limited by data heterogeneity, as certain domains are not presented in our dataset. Fourth, our best-performing LLM, mT5 still struggles with long-form Bangla text which leads to a potential loss of contextual details.

References

Muskan Ahmed, ASM Siddiqui, and Ayon Das. 2024. *Enhancing Bangla text summarization in a monolingual setting*. Ph.D. thesis, Brac University.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021a. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *arXiv preprint arXiv:2205.11081*.

Prithwiraj Bhattacharjee, Avi Mallick, and Md Saiful Islam. 2021b. Bengali abstractive news summarization (bans): a neural attention approach. In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*, pages 41–51. Springer.

Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpbdpatriots at blp-2023 task 2: A transfer learning approach towards bangla sentiment analysis. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 286–292.

SM Tasnimul Hasan, Md Ashfaqur Rahman, Md Mahamudul Hasan, Mohammad Rakibul Hasan, and Md Moinul Hoque. 2023. Advancing abstractive bangla text summarization: A deep learning approach using seq2seq encoder-decoder model and t5 transformer. In *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pages 1–6. IEEE.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

SM Afif Ibne Hayat, Avishek Das, and Mohammed Moshiul Hoque. 2023. Abstractive bengali text summarization using transformer-based learning. In *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6. IEEE.

Alam Khan, Sanjida Akter Ishita, Fariha Zaman, Ashiqul Islam Ashik, and Md Moinul Hoque. 2023. Intelligent combination of approaches towards improved bangla text summarization. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pages 1–6. IEEE.

Asif Ahammad Miazee, Tonmoy Roy, Md Robiul Islam, and Yeamin Safat. 2025. Abstractive text summarization for bangla language using nlp and machine learning approaches. *arXiv preprint arXiv:2501.15051*.

Aditya Mukherjee. 2022. *Developing Bengali Text Summarization with Transformer Base model*. Ph.D. thesis, Dublin, National College of Ireland.

Chester Palen-Michel and Constantine Lignos. 2022. Lr-sum: Summarization for less-resourced languages. *arXiv preprint arXiv:2212.09674*.

Mohammad Abu Tareq Rony and Mohammad Shariful Islam. 2024. Evaluating large language models for summarizing bangla texts. In *Eighth Widening NLP Workshop (WiNLP 2024) Phase II*.

Anusree Roy. 2024. midnightglow/banglasum · datasets at hugging face — huggingface.co. https://huggingface.co/datasets/midnightGlow/BanglaSum. The dataset was created by web scraping different online newspapers like 'The Daily Star', 'Prothom Alo', and 'BBC News Bangla' using the Beautiful Soup library of Python.

GM Shahariar, Tonmoy Talukder, Rafin Alam Khan Sotez, and Md Tanvir Rouf Shawon. 2023. Rank your summaries: Enhancing bengali text summarization via ranking-based approach. In *International Conference on Big Data, IoT and Machine Learning*, pages 153–167. Springer.

Anupam Singha and NR Rajalakshmi. 2023. Bengali text summarization with attention-based deep learning. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–5. IEEE.

Mariyam Sultana, Partha Chakraborty, and Tanupriya Choudhury. 2022. Bengali abstractive news summarization using seq2seq learning with attention. In *Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021*, pages 279–289. Springer.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Quantitative Analysis with Existing Datasets

This appendix presents Table 2 showing the quantitative analysis of our proposed *BWSD* with the existing state-of-the-art datasets.

B Hyper-parameter Settings

Since Bangla is a complex language with unique grammatical structures and rich contextual dependencies, proper tuning ensures the model can effectively capture these features. Moreover, the optimized hyper-parameters increase the efficiency of training and reduce the computational cost, hence making summarization both more accurate and resource-effective. Table 3 presents the hyper-parameter setting of our proposed BATSS.

C Comparative Analysis with Existing Datasets

This appendix presents Table 4 showing the evaluation score of *BWSD* and existing datasets using BATSS.

Dataset	L.	# of data	W.C.	U.W.C.	Length of Articles	Length of Summaries
XL_Sum (Hasan et al., 2021)	M	10126	1.2 million	155267	400-1800	20-120
BANS (Bhattacharjee et al., 2021b)	B	19096	857341	44318	05-76	03-12
BanglaSum (Roy, 2024)	B	9311	6.5 million	185376	14-3500	01-1428
LR_Sum (Palen-Michel and Lignos, 2022)	M	715	27288	10583	150-450	25-45
BWSD	B	1100	283573	23979	100-1400	25-135

Table 2: Comparative analysis of existing datasets (L. - Language [M: Multilingual; B: Bangla], W.C.- Word Count, U.W.C. - Unique Word Count)

Parameter	Value
Evaluation strategy	epoch
Learning rate	5e-5
Train batch size	8
Evaluation batch size	8
Number of train epochs	40
Weight decay	0.01
Optimizer	Adam

Table 3: Hyper-parameter settings for proposed BATSS

Ref.	ROUGE Score								
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
	BanglaT5			BanglaBERT			mT5		
BANS (Sultana et al., 2022), (Mukherjee, 2022), (Miazee et al., 2025)	15.32	2.81	15.08	17.88	6.49	14.97	22.14	5.16	21.29
XL_Sum (Hasan et al., 2021), (Shahariar et al., 2023), (Ahmed et al., 2024)	16.55	5.88	14.24	25.36	7.02	16.78	27.86	10.91	24.85
BANS + XL_Sum (Hayat et al., 2023)	11.25	3.34	9.88	21.38	7.26	16.20	19.46	7.8	17.73
BTS (Singha and Rajalakshmi, 2023)	17.16	7.66	14.06	24.05	9.18	18.70	30.18	13.92	24.36
BNLPC (Hasan et al., 2023), (Rony and Islam, 2024), (Khan et al., 2023)	19.77	9.52	13.49	38.43	**17.21**	32.93	32.57	14.42	23.39
Proposed *BWSD* dataset	25.36	**12.97**	21.09	38.01	16.95	31.76	41.68	**22.57**	32.28

Table 4: Comparative study with existing works