BanglaClickBERT: Bangla Clickbait Detection from News Headlines using Domain Adaptive BanglaBERT and MLP Techniques

Saman Sarker Joy, Tanusree Das Aishi, Naima Tahsin Nodi, Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University, 66 Mohakhali, Dhaka-1212, Bangladesh
{saman.sarker.joy, tanusree.das.aishi, naima.tahsin.nodi}@g.bracu.ac.bd
annajiat@bracu.ac.bd

Abstract

News headlines or titles that deliberately persuade readers to view a particular online content are referred to as clickbait. There have been numerous studies focused on clickbait detection in English language, compared to that, there have been very few researches carried out that address clickbait detection in Bangla news headlines. In this study, we have experimented with several distinctive transformers models, namely BanglaBERT and XLM-RoBERTa. Additionally, we introduced a domain-adaptive pretrained model, BanglaClickBERT. We conducted a series of experiments to identify the most effective model. The dataset we used for this study contained 15,056 labeled and 65,406 unlabeled news headlines; in addition to that, we have collected more unlabeled Bangla news headlines by scraping clickbait-dense websites making a total of 1 million unlabeled news headlines in order to make our BanglaClickBERT. Our approach has successfully surpassed the performance of existing state-of-the-art technologies providing a more accurate and efficient solution for detecting clickbait in Bangla news headlines, with potential implications for improving online content quality and user experience.

1 Introduction

The Internet has led to a surge in the use of online news media, which provides users with easy access to information at any time. However, news websites use clickbait headlines that can be misleading and frustrating to users. These headlines are designed to attract users and create suspense, containing exaggerated information that does not match the content. Clickbait headlines aim to lure users into clicking on them but ultimately cause frustration. Pengnate et al. concluded a research and found that clickbait headlines can lead to higher click-through rates, but may lead to negative user experiences such as frustration and disappointment. Examples of clickbait headlines in Bangla are in Figure 1. The core differences between clickbait and non-clickbait is described in Appendix A.

Figure 1: Examples of Bangla clickbait news headlines with its corresponding English translation and type of clickbait

The use of online news media has increased rapidly in Bangladesh, with an estimated 66.3 million internet users[1] and 14 million online readers of Prothom Alo (Correspondent, 2022), one of the top newspapers in the country. However, the increasing number of clickbait titles on news websites has become a significant issue, leading to frustration and disappointment among users. While research has been conducted on clickbait detection in English, very little has been done in Bangla, a language spoken by millions of people in Bangladesh and other countries. In English, for The Clickbait Challenge 2017, Webis Clickbait Corpus 2017 (Potthast et al., 2018b) was created which had a total of 38,517 sentences from major US news publishers. In Bangla, Mahtab et al. have constructed a Bangla clickbait detection dataset containing 15,056 labeled news articles and 65,406 unlabelled news articles. In this paper, we present BanglaClickBERT, a pretrained model for clickbait detection in

[1] https://www.cia.gov/the-world-factbook/countries/bangladesh/

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 1–10
November 29 - December 1, 2023

Bangla news websites. We use the labeled dataset for training and validating our model and scrape clickbait-dense websites to gather more unlabelled news article headlines, increasing the number of unlabelled news headlines to around 1 million. We use this to pretrain the BanglaBERT (Bhattacharjee et al., 2022) model, which we then pretrain to create BanglaClickBERT.

The main contributions of this paper can be summarized as follows:

- We scrape clickbait-dense websites and create an unlabelled news headlines dataset of around 1 million which we use to pretrain BanglaBERT model converting it to BanglaClickBERT.
- We experiment with different machine learning models, deep neural network models, and transformers models like BanglaBERT, XLM-RoBERTa, and our BanglaClickBERT to develop a Bangla Clickbait Detection model for Bangla news headline data. We compare the performance of our model using different metrics.

2 Literature Review

The roots of clickbait can be found in tabloids, a form of journalism that has existed since the 1980s (Bird, 2008). The three primary sources from which clickbait identification attributes may be generally retrieved are (1) the related article that the post text wants the user to visit, (2) metadata for both, and (3) the connected article (Munna and Hossen, 2021). Potthast et al. and Biyani et al. additionally took into account metadata, related content, and handcrafted elements in addition to the post-text analysis. They used methods like Gradient Boosted Decision Trees (GBDT) and assessed the TF-IDF similarity between the headline and article content. Potthast et al. in another paper also mentioned the Clickbait Challenge 2017, which invited the affirmation of 13 detectors were presented as the clickbait detectors for screening, realizing considerable enhancements in detecting performance above the prior state of the art. Zhou first used a self-attentive RNN to choose the crucial terms in the title before building a BiGRU network to encode the contextual information for the 2017 Clickbait Challenge. On the contrary, Thomas used an LSTM model for the clickbait challenge that included article content. To create the word embedding of clickbait titles, Rony et al. applied the continuous skip-gram model. Nevertheless, Indurthi et al. were the first to study the use of transformer regression models in clickbait identification and won the clickbait challenge. Additionally, Hossain et al. produced the first dataset of Bengali newspapers for Bengali false news detection of around 50K Bangla news articles in an annotated dataset. Besides Bangla, we have explored about clickbait detection techniques in news and social media in other languages. Genç and Surer used Logistic Regression (85% accuracy), Random Forest (86% accuracy), LSTM (93% accuracy), ANN (93% accuracy), Ensemble Classifier (93% accuracy), and BiLSTM (97% accuracy) on 48,060 headlines from news sources pulled from Twitter for Turkish clickbait detection. Moreover, Razaque et al. used Long short-term memory, Word2vec and compared their models with Naive Bayes classifier for clickbait detection on social media. Bronakowski et al. achieved 98% accuracy in recognizing clickbait headlines by using thirty distinct types of semantic analysis and six different machine-learning approaches, both individually and in groups. The suggested models can be used as a model for creating useful programs that swiftly identify clickbait headlines. Farhan et al. used Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN)-based ensemble model for sarcasm detection for Bangla language achieving 96% F1-score and accuracy. It gave us an insight on what type of work can be done using NLP and for gathering knowledge and examples related to our work. Additionally, Beltagy et al. created SciBERT which is a pretrained language model, based on BERT used unsupervised pretraining on scientific articles, providing us knowledge about domain-adaptive BERT which can help enhance efficiency on a range of scientific NLP tasks and produce cutting-edge results. Moreover, Jahan et al. created BanglaHateBERT, which is a retrained version of the pre-existing BanglaBERT model, and trained it having a widespread corpus of hostile, insulting, and offensive Bengali language, and outperformed the generic pretrained language model in various datasets. So, to sum up with, we have analysed about the origins of clickbait, checked different datasets on different languages, learned about differnet NLP methods, and observed the potentials of specialized transformers models like SciBERT and BanglaHateBERT.

3 Problem Statement

We approach the task of clickbait detection as a decision-making challenge; a binary classification task problem with two main categories $C = \{clickbait, non - clickbait\}$. Given a set of Bangla news headlines $T = \{t_1, t_2, t_3, \ldots, t_N\}$, our objective is to predict labels $Y = \{y_1, y_2, y_3, \ldots, y_n\}$ for these headlines. Here, y_i assumes the value 1 if title t_i is classified as clickbait and 0 if it is classified as non-clickbait. The problem can be formulated as,

$$< C, Y >= \{non - clickbait : 0, clickbait : 1\}$$

4 Dataset Description

The dataset (Mahtab et al., 2023) we used consists of two sets: an annotated set and an unannotated set of clickbait news information. The information with our augmentation is shown in Table 1.

4.1 Annotated Dataset

The annotated dataset comprises 15,056 articles, each labeled with one of two categories: Clickbait as 1 and Non-clickbait as 0. The articles in this subset cover a diverse range of topics. For our task, we focus only on the columns "Headlines" and "Labels" as they are essential. This dataset will be used for the classification task.

4.2 Unannotated Dataset

The unannotated dataset consists of 65,406 Bangla articles with clickbait titles. These articles were gathered from clickbait-dense websites. However, since 65k unlabelled samples may not be sufficient for our task, we expanded the dataset by scraping more clickbait-dense websites using *Selenium*[2] library. This effort resulted in a total of 1,078k or 1 Million unlabelled clickbait headlines. This unannotated dataset will be used for the pretraining.

Information	Value
Crawling Period	Feb 2019 - June 2023
Total Clickbait	5,239
Total Non-clickbait	9,817
Total Unlabelled Before	65,406
Total Unlabelled After	1,078,234

Table 1: Information of both the annotated and unannotated datasets

[2] https://www.selenium.dev/

5 Methodology

We have used some Statistical Models and Deep Learning Models and then we have implemented Transformers models Like BanglaBERT, XLM-RoBERTa and Domain Adaptive BanglaClickBERT with several variations. Based on these variation, we try to come up with the best model.

5.1 Statistical Models

For statistical models, we will employ Logistic and Random Forest classifiers on a combination of various features like TF-IDF (term frequency–inverse document frequency) of the word and character n-grams, Bangla pretrained word embeddings, punctuation frequency, and normalized Parts-of-Speech frequency.

5.2 Deep Learning Models

When it comes to deep learning models, there are several powerful techniques that can be employed e.g. Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM) and ensemble methods. These models have shown great success in various natural language processing tasks, including sentiment analysis and text classification.

5.3 Transformer Models

5.3.1 BanglaBERT

BanglaBERT (Bhattacharjee et al., 2022) is a BERT-based Natural Language Understanding (NLU) model pretrained specifically on Bangla using a massive 27.5GB pretraining corpus. BanglaBERT has demonstrated remarkable performance in achieving state-of-the-art results across diverse NLP tasks.

5.3.2 XLM-RoBERTa

XLM-RoBERTa (Conneau et al., 2020), a large-scale multilingual language model based on Facebook's RoBERTa (Liu et al., 2019). XLM-RoBERTa undergoes pretraining on an extensive 2.5TB dataset of filtered CommonCrawl data.

5.3.3 Domain Adaptive Pretraining

We also propose to further pretrain BanglaBERT using a large number of headlines extracted from clickbait-filled websites. Gururangan et al. finds that tailoring pretrained language models to specific domains through adaptive pretraining techniques leads to significant improvements in task performance.

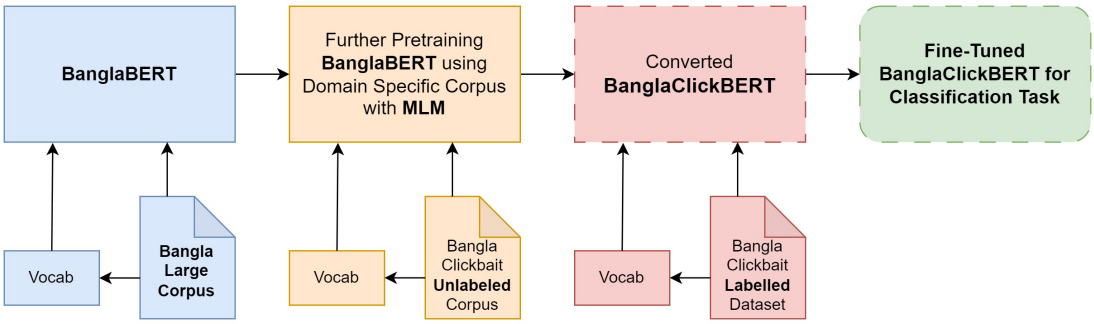

Figure 2: Workflow of *BanglaClickBERT* Creation

6 Creation of BanglaClickBERT

Language models like BERT have revolutionized the field of NLP by introducing context-aware learning and significantly improving performance across various NLU tasks. However, applying these models to low-resource languages such as Bangla requires specialized adaptation to achieve optimal results. To address this challenge, we propose the development of BanglaClickBERT by further pretraining BanglaBERT with a vast dataset of clickbait news headlines. A workflow of this is shown in Figure 2.

6.1 Reason for Pretraining

Gururangan et al. investigated whether it is still helpful to tailor a pretrained model to the domain of a target task. From their research, it was found that a second phase of pretraining in-domain (domain-adaptive pretraining) leads to performance gains, in both high and low-resource settings. Also, in the BanglaHateBERT paper (Jahan et al., 2022), we found performance gains after pretraining.

6.2 Pretraining Data

We collected a diverse set of clickbait news headlines mentioned in Section 4, comprising 1 million samples from various online sources. These headlines were chosen to cover a wide range of clickbait headlines, ensuring the model's adaptability to different contexts like news on lifestyle, entertainment, business, viral videos etc.

6.3 Training Strategy

The retraining process was carried out using the Masked Language Model (MLM) approach. During training, we masked 15% of the tokens in each sequence, forcing the model to predict these masked tokens and thus gain contextual understanding. Additionally, we set the model to accept up to 128 sentence tokens to capture more extensive contextual dependencies. BanglaClickBERT was pretrained for 10 epochs, on an *NVIDIA GeForce RTX 3070*. It took us almost 28 hours to pretrain for 10 epochs. We adopted the Adamw (Loshchilov and Hutter, 2019) optimization solver, known for its computational efficiency and memory-friendly characteristics, with a learning rate of 5e-5. The maximum sequence length was set to 32 as there was no sentence bigger than 30 shown in Figure 3. The pretrained models are uploaded on Hugging face website.[3]. The unannotated dataset of clickbaits will also be provided on request.[4].

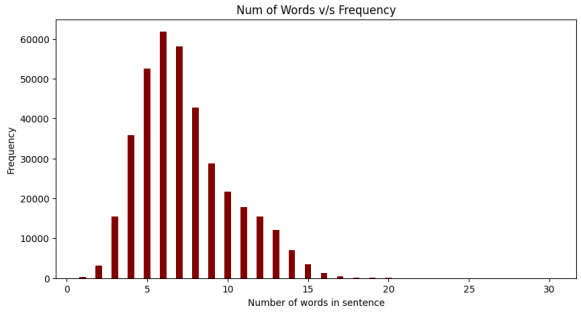

Figure 3: Frequnecy of all the sentences in the unannotated corpus. It shows that all the sentence lengths are less than 30.

7 System Overview

7.1 Statistical Models

We used two Statistical models: Logistic Regression and Random Forest. Logistic Regression and Random Forest both are widely used classification algorithms that are particularly well-suited for binary classification tasks. We used TF-IDF vectors.

[3] https://huggingface.co/samanjoy2/banglaclickbert_base

[4] https://tinyurl.com/BanglaClickBERTdata

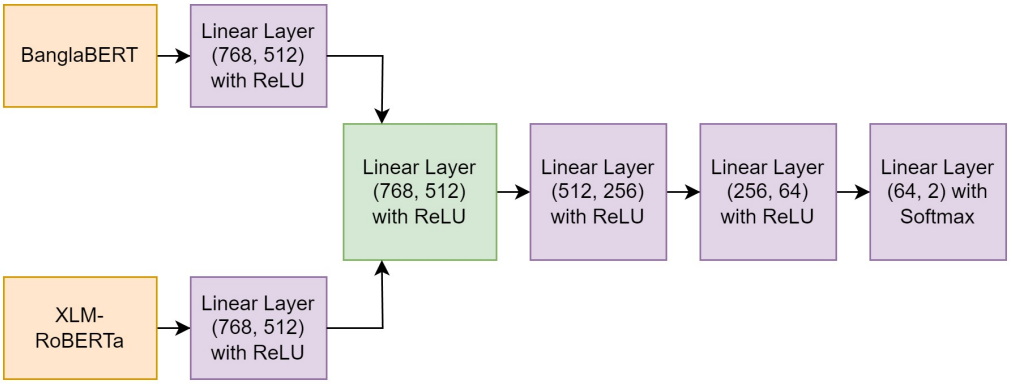

Figure 4: BanglaBERT and XLM-RoBERTa concatenation of the last layer + MLP Architecture

This captures the sequential patterns of characters in the text using character n-grams of lengths 1, 2, 3, 4 and 5. For example, for n=3, the word "hello" would be represented as [hel, ell, llo]. These character n-grams can capture important linguistic information and patterns in the text, such as common prefixes, suffixes, and other recurring character sequences.

7.2 Deep Learning Models

We used two deep learning models: the Bi-LSTM Network model and Ensemble of Convolutional neural network + Gated recurrent unit (Farhan et al., 2023) both with Bengali GloVe Embeddings (Sarker, 2021). Bengali GloVe Pretrained Word Vectors was pretrained with Wikipedia and crawled news articles with 39 million tokens and has a 0.18 million vocab size. We used the 300d vector version.

7.3 Transformer Models

Throughout our experimentation, we have explored various architectural configurations for these transformer models. To illustrate the general architecture that we employed, we present an example in Figure 4.

7.3.1 BanglaBERT / XLM-RoBERTa / BanglaClickBERT (last layer) + MLP

In this setup, the last layer of the BanglaBERT and XLM-RoBERTa base models are used as the input. The last layer contains contextualized information learned from pretraining on the Bangla and multilingual data, respectively. These representations are then passed through additional linear layers and fine-tuned on the specific task or dataset during the training phase. This allows the model to adapt to the task while benefiting from the pretrained language representation capabilities of BanglaBERT and XLM-RoBERTa.

7.3.2 BanglaBERT / XLM-RoBERTa / BanglaClickBERT (average of all layers) + MLP

Instead of using only the last layer, this setup takes the average of all layers in the BanglaBERT and XLM-RoBERTa base models. By doing so, the model can incorporate information from various depths of the transformers, capturing different levels of context and features. The averaged representations are then fed into linear layers and fine-tuned for the specific task.

7.3.3 BanglaBERT / BanglaClickBERT and XLM-RoBERTa concatenation of the last layer + MLP

In this approach, the outputs from the last layers of BanglaBERT and XLM-RoBERTa are concatenated together. This allows the model to combine the representations learned by each transformer independently. The concatenated representations are then fed into an MLP (multi-layer perceptron) with fully connected layers before producing the final output, which is the prediction for the given task.

8 Experimental Setup

8.1 Prepossessing

The dataset already underwent comprehensive preprocessing, removing HTML tags, URL links, newline escape sequences and emojis. They also preserved all syntactically correct punctuation in the titles and removed punctuation that appeared in the middle of words.

SL	Model Names	Precision	Recall	F1-Score	Accuracy
1	Logistic Regression (with TF-IDF 1-5 n-grams)	0.6540	0.3745	0.4763	0.7102
2	Random Forest (with TF-IDF 1-5 n-grams)	0.6789	0.4509	0.5419	0.7317
3	Bi-LSTM Network (with GloVe Embeddings)	0.6544	0.5877	0.6192	0.7457
4	Ensemble of CNN + GRU (with GloVe Embeddings) (Farhan et al., 2023)	0.6774	0.6103	0.6421	0.7606
5	GAN-BanglaBERT (Mahtab et al., 2023)	0.7545	0.7481	0.7512	**0.8257**
6	BanglaBERT last layer + MLP	0.7377	0.7241	0.7308	0.8088
7	BanglaBERT Large last layer + MLP	0.7349	0.7328	0.7338	0.8124
8	XLM-RoBERTa last layer + MLP	0.7038	**0.7505**	0.7264	0.8134
9	Domain Adaptive BanglaClickBERT last layer + MLP	0.7802	0.7081	0.7424	0.8094
10	BanglaBERT avg of all layers + MLP	0.7293	0.7138	0.7214	0.8018
11	XLM-RoBERTa avg of all layers + MLP	0.6962	0.6474	0.6709	0.7596
12	Domain Adaptive BanglaClickBERT avg of all layers + MLP	0.7717	0.7343	0.7525	0.8214
13	BanglaBERT + XLM-RoBERTa + Embeddings concatenated. Before concatenating passed through one linear layer. Followed by MLP	0.7821	0.7153	0.7472	0.8138
14	Domain Adaptive BanglaClickBERT + XLM-RoBERTa + Embeddings concatenated. Before concatenating passed through one linear layer. Followed by MLP	**0.7896**	0.7234	**0.7551**	0.8197

Table 2: Performance comparison of different Models. Precision, Recall and F1-Score are for the *clickbait* class.

We, furthermore, for our research, extended the preprocessing paradigm by using the Abugida Normalizer and Parser for Unicode Texts (bnunicodenormalizer)[5], enhancing the overall data quality and compatibility. This advanced technique played a pivotal role in fine-tuning later on.

8.2 Experimental Settings

We will be using Statistical models and Deep learning models for Baseline Creation. Then we will be using transformer models. Our main focus is on using Transformer. We have used Transformers with several variations. Based on this variation we try to come up with the best model. For the statistical models, we used TF-IDF vectors and n-grams length from 1 to 5. For the deep learning models, we used 300d Bangla GloVe embeddings. We used a variation of transformers models which we described earlier. We have chosen to mainly use the base (12 layers) versions of these models, as the large (24 layers) models will be computationally expensive and unnecessary for our task. We experimented with BanglaBERT Large model, however, it was providing similar results (discussed in section 9) to the BanglaBERT base model. So, for further experimentation, we continued with the base models. For hyperparameters, we have taken the number of epochs for training as 20, the learning rate is 1e-5, maximum length is 32, batch size of 128, the loss function is Cross Entropy Loss and the optimizer is AdamW (Loshchilov and Hutter, 2019) in all the models. The labeled dataset is divided into three distinct subsets: the training set, test set, and validation set. This allocation was thoughtfully proportioned, with 70% (10839 headlines) of the data reserved for training, 20% (3012 headlines) for testing, and 10% (1205 headlines) for validation purposes. We used the same data splits used in (Mahtab et al., 2023) that helps us to compare with this technique properly. We have used the precision, recall, macro F1-Score and accuracy as measures of evaluation.

[5] https://pypi.org/project/bnunicodenormalizer/

9 Results and Analysis

As depicted in Table 2, the statistical models, namely Logistic Regression and Random Forest, failed to identify the clickbait articles fruitfully and exhibited unsatisfactory performance. The deep learning model; The Bi-LSTM model achieved an F1-score of 61.92%. The Ensemble of CNN + GRU (Farhan et al., 2023) performed even better with an F1-score of 64.21%. This highlights the advantages of using word embeddings and sequence modeling for clickbait detection in Bangla. However, there is still considerable room for improvement, as the overall F1-scores remained relatively low.

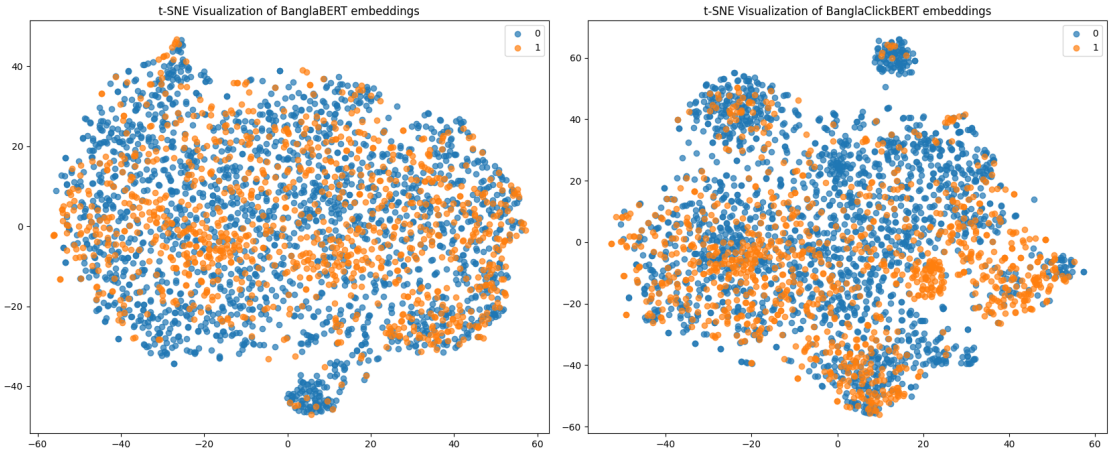

Figure 5: Visualization of last layer hidden representations using t-SNE (van der Maaten and Hinton, 2008) for BanglaBERT (Left) and BanglaClickBERT (Right) without any fine-tuning. 0 represents *non-clickbait* and 1 represents *clickbait* in both figures.

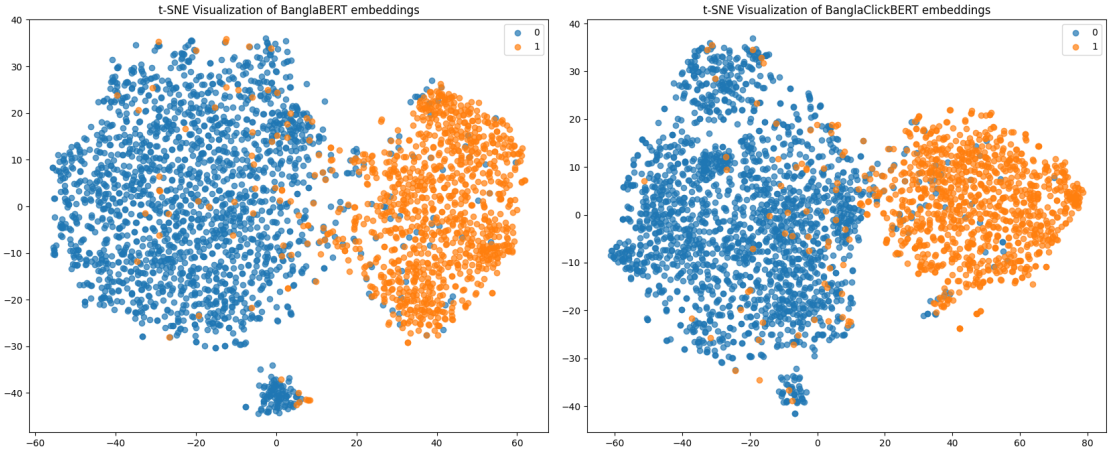

Figure 6: Visualization of last layer hidden representations using t-SNE (van der Maaten and Hinton, 2008) for BanglaBERT (Left) and BanglaClickBERT (Right) with fine-tuning. 0 represents *non-clickbait* and 1 represents *clickbait* in both figures.

We then from (Mahtab et al., 2023) paper found that, their approach GAN-BanglaBERT achieved a 82.57% accuracy which is the highest accuracy among all the models we described or experimented with. However, this high accuracy does not give us the whole picture as the dataset is imbalanced and a proper evaluation should be according to the macro F1-score which in its case is 75.12% for clickbait class.

Transformer models; BanglaBERT, XLM-RoBERTa and BanglaClickBERT demonstrated consistent improvements over all other approaches. In particular, using only the last layer in conjunction with MLP yielded excellent results. Notably, both BanglaBERT base and BanglaBERT Large performed similarly, 73.08% and 73.38%, indicating that increasing the model's parameters did not contribute significantly to this specific clickbait detection task. On the other hand, the Domain Adaptive BanglaClickBERT exhibited better than that of BanglaBERT and XLM-RoBERTa, respectively, with F1 score of 74.24%. This underscores the effectiveness of pretraining the model with domain-adaptive data for clickbait detection.

Considering the average of all layers proved to be both advantageous and disadvantageous as it captured more informative representations. The F1-score of BanglaBERT and XLM-RoBERTa decreased, whereas, it proved to be beneficial for Domain Adaptive BanglaClickBERT as taking the average of all its layer increased the F1-score by 1.01%. It again proves that pretraining the layers of BanglaBERT has helped all its layers to understand more about clickbait sentences.

Model Names	Attention Weighted Words
BanglaBERT	[CLS] এক মিস ##কল ##লেই মধুর সম্পর্ক সর্বনাশ তরুণীর [UNK] . দেখুন (ভিডিও) [SEP]
BanglaClickBERT	[CLS] এক মিস ##কল ##লেই মধুর সম্পর্ক সর্বনাশ তরুণীর [UNK] . দেখুন (ভিডিও) [SEP]
Raw Headline	এক মিস কলেই মধুর সম্পর্ক সর্বনাশ তরুণীর. দেখুন (ভিডিও)
Translated Headline	One missed call destroys the sweet relationship of the young woman. Watch (Video)

Table 3: Comparison between finetuned BanglaBERT and finetuned BanglaClickBERT. A clickbait sentence is chosen and both the model predict it as clickbait. Each word is highlighted according to their attention weights.

Moreover, concatenating the embeddings from two pretrained language models further enhanced performance, illustrating that combining related models could capture complementary information for clickbait detection in Bangla. The combination of Domain Adaptive BanglaClickBERT and XLM-RoBERTa achieved the highest F1-score of 75.51% for the clickbait class surpassing other models we discussed about including the GAN-BanglaBERT (Mahtab et al., 2023).

In terms of precision, recall, F1-score, and accuracy, the Domain Adaptive BanglaClickBERT model proved to be more consistent to all other models. However, since the F1-scores were tightly clustered within the range of 0.74 to 0.75, to support our claims, we ran each model ten times with different seeds and conducted a statistical test. The model, labeled *"Domain Adaptive BanglaClickBERT + XLM-RoBERTa + Embeddings concatenated. Before concatenating passed through one linear layer. Followed by MLP"* outperforms all other models, and the difference in performance is statistically significant ($p < 0.05$) according to McNemar's test (Dietterich, 1998)."

This finding is further supported by the t-SNE visualization depicted in Figure 5 and Figure 6. The t-SNE visualization effectively shows how these models, even without fine-tuning of the training data, group their predictions. It becomes evident that BanglaClickBERT exhibits better clustering than BanglaBERT, underscoring the idea that training BanglaClickBERT can enhance the learned representations and subsequently improve overall performance. This can be shown more prominently in Figure 6 that BanglaClickBERT managed to cluster the embeddings of clickbait headlines better than BanglaBERT.

Additionally, as illustrated in Table 3, using the *Transformers Interpret* (Pierse, 2021) we tried to analyse how the models predict their predictions. Green highlights indicate supportive words for the prediction, while red highlights show opposing words. Brightness reflects the strength of their contribution or opposition. We can see that, finetuning the BanglaBERT and BanglaClickBERT models results in different attention patterns for words. In particular, BanglaClickBERT allocates greater attention to words related to clickbait, a characteristic that BanglaBERT does not achieve.

In conclusion, the results suggest that BanglaClickBERT, proves to be highly effective for clickbait detection in Bangla. If more and better labelled data is used to finetune this, this approach will perform better than other approaches.

10 Conclusion

In conclusion, this study represents a significant advancement in the field of clickbait detection, particularly for the Bangla language, where research has been limited. While clickbait detection in English has been extensively studied, the Bangla news headlines have been largely overlooked. To address this gap, we conducted a comprehensive analysis using state-of-the-art transformer models, such as BanglaBERT, XLM-RoBERTa, and the newly developed BanglaClickBERT. We enhanced the performance of these models by incorporating MLP methods to achieve the best results. To bolster the research, we augmented the dataset by including an additional 1 million unlabeled Bangla news headlines, sourced from clickbait-dense websites. This expanded dataset significantly empowered the BanglaClickBERT model. Through rigorous experimentation and testing, our approach showed better results compared to existing state-of-the-art techniques. Our work not only contributes to the improvement of clickbait identification in Bangla news headlines but also fills the void in research in this language domain. As clickbait continues to impact the way information is consumed, our findings will be valuable for media organizations, content creators, and platforms to promote responsible and reliable information dissemination in the Bangla-speaking community.

11 Acknowledgement

We extend our sincere appreciation to the dedicated faculty members of the Computer Science and Engineering department at BRAC University, Bangladesh. Their unwavering support and valuable guidance have been integral to our research efforts. Additionally, we express our gratitude to the anonymous reviewers whose constructive feedback has significantly contributed to the improvement of our research work.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

S. Elizabeth Bird. 2008. *Tabloidization*. John Wiley & Sons, Ltd.

Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *AAAI Conference on Artificial Intelligence*.

Mark Bronakowski, Mahmood Al-khassaweneh, and Ali Al Bataineh. 2023. Automatic detection of clickbait headlines using semantic analysis and machine learning techniques. *Applied Sciences*, 13(4).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Staff Correspondent. 2022. Prothom alo at the top with 5 million readers. *Prothomalo*. [Online].

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*, 10(7):1895–1923.

Niloy Farhan, Ishrat Tasnim Awishi, Md Humaion Kabir Mehedi, MD. Mustakin Alam, and Annajiat Alim Rasel. 2023. Ensemble of gated recurrent unit and convolutional neural network for sarcasm detection in bangla. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0624–0629.

Sura Genç and Elif Surer. 2021. Clickbaittr: Dataset for clickbait detection from turkish news sites and social media with a comparative analysis via machine learning algorithms. *Journal of Information Science*, 49:480 – 499.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. BanFakeNews: A dataset for detecting fake news in Bangla. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France. European Language Resources Association.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Gupta, and Vasudeva Varma. 2020. Predicting clickbait strength in online social media. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4835–4846, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022. BanglaHateBERT: BERT for abusive language detection in Bengali. In *Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis*, pages 8–15, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Md. Motahar Mahtab, Monirul Haque, and Farig Sadeque. 2023. Banglabait: Semi-supervised adversarial approach for clickbait detection on bangla clickbait dataset. In *Proceedings of Recent Advances in Natural Language Processing*, pages 744–754, Varna, Bulgaria.

Mahmud Hasan Munna and Md Shakhawat Hossen. 2021. Identification of clickbait in video sharing

platforms. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pages 1–6.

Supavich Fone Pengnate, Jeffrey Chen, and Alex Young. 2021. Effects of clickbait headlines on user responses: An empirical investigation. *Journal of International Technology and Information Management*, 30(3):1.

Charles Pierse. 2021. Transformers Interpret.

Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. 2018a. The clickbait challenge 2017: Towards a regression model for clickbait strength. *CoRR*, abs/1812.10847.

Martin Potthast, Tim Gollub, Matti Wiegmann, Benno Stein, Matthias Hagen, Kristof Komlossy, Sebstian Schuster, and Erika P. Garces Fernandez. 2018b. Webis clickbait corpus 2017 (webis-clickbait-17).

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *Advances in Information Retrieval*, pages 810–817, Cham. Springer International Publishing.

Abdul Razaque, Bandar Alotaibi, Munif Alotaibi, Shujaat Hussain, Aziz Alotaibi, and Vladimir Jotsov. 2022. Clickbait detection using deep recurrent neural network. *Applied Sciences*, 12(1).

Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 232–239, New York, NY, USA. Association for Computing Machinery.

Sagor Sarker. 2021. BNLP: natural language processing toolkit for bengali language. *CoRR*, abs/2102.00405.

Philippe Thomas. 2017. Clickbait identification using neural networks.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Yiwei Zhou. 2017. Clickbait detection in tweets using self-attentive network.

A Appendix

A.1 Distinguishing between clickbait and non-clickbait

The following examples shows the differences between clickbait headlines and non-clickbait headlines:

Examples 1:

নায়কের চেয়েও বেশি টাকা আয় করেন
এই কমেডিয়ান ! জানেন কত... ?

Translated: This comedian earns more money than the hero! Do you know how much...?

Examples 2:

দীপিকার ব্যাগে যা থাকে! (ভিডিও)

Translated: What's in Deepika's bag! (video)

Examples 3:

বলিউডের ফিল্মফেয়ার অ্যাওয়ার্ড পেতে
যাচ্ছেন বাংলাদেশের তন্বী

Translated: Bangladesh's Tanvi is going to receive Bollywood's Filmfare Award

In the above mentioned examples we can observe that clickbait headlines have distinguish patterns. In the Example 1, the headline does not immediately reveal who the more-earning comedian is. Instead, it keeps the reader in suspense, prompting them to click in order to satisfy their curiosity and uncover the answer.

Example 2, the headline uses the "Curiosity Gap" technique by teasing an intriguing element of the story without giving away the full details. Moreover, the excessive use of punctuation or other symbols, such as exclamation points, is often used to heighten the reader's curiosity and create a sense of urgency or excitement.

On the other hand, Example 3 falls into the category of non-clickbait due to its straightforward and informative headline. It effectively communicates the content and purpose of the article, leaving no room for curiosity or teasing. This type of headline is transparent and does not rely on sensationalism or misleading tactics to attract readers, making it a clear example of non-clickbait.