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

Saman Sarker Joy*, Tanusree Das Aishi†, Naima Tahsin Nodi‡, Md Sabbir Hossain§ & Annajiat Alim Rasel¶
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University
66 Mohakhali, Dhaka - 1212, Bangladesh
*saman.sarker.joy@g.bracu.ac.bd, †tanusree.das.aishi@g.bracu.ac.bd, ‡naima.tahsin.nodi@g.bracu.ac.bd,
§ext.sabbir.hossain@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, XLMRoberTa and we will make a Domain-specific BanglaBERT named BanglaClickBERT and each of them along with MLP techniques, in order to come up with the best performing model. The dataset we used for this study contained 15,056 headlines with labels and 65,406 unlabeled news headlines, in addition to that we have collected more unlabeled Bangla news headlines around 400k in order to make our BanglaClickBERT. We expect that our approach will work effectively to deal with the task of Bangla clickbait detection.**

***Index Terms*—Clickbait, BanglaBERT, MLP**

I. 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, some news websites use clickbait headlines that can be misleading and frustrating for users. These headlines are designed to attract users and create suspense, often containing exaggerated information that does not match the actual content. Clickbait headlines aim to lure users into clicking on them but ultimately cause frustration. Study [1], found that clickbait headlines can lead to higher click-through rates, but may also lead to negative user experiences such as frustration and disappointment.

The use of online news media has increased rapidly in Bangladesh, with an estimated 66.3 million internet users [2] and 14 million online readers of Prothom Alo [3], 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,

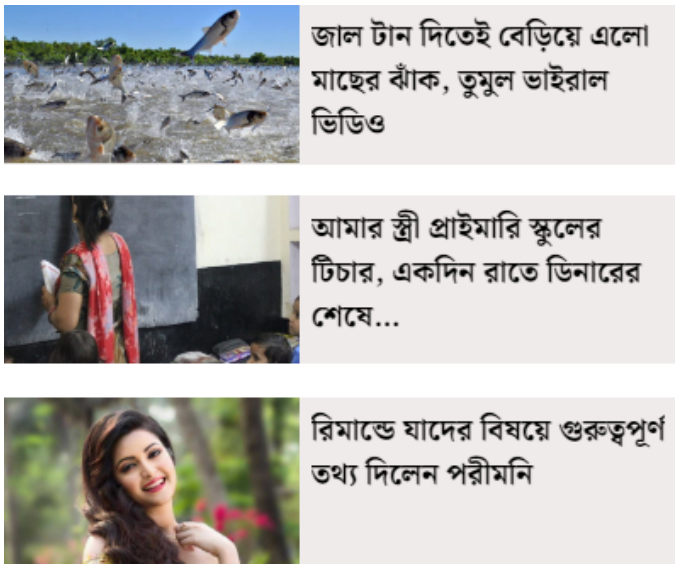

Fig. 1. Examples of Bangla clickbait news titles.

Webis Clickbait Corpus 2017 [4] was created which had a total of 38,517 sentences from major US news publishers. In Bangla, BanglaBait [5], where they 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 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 400k. We use this to pretrain the BanglaBERT [6] model, which we then fine-tune 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 400k to pretrain

our BanglaBERT model, which we then fine-tune to create BanglaClickBERT, a pretrained model for clickbait detection in Bangla news websites.
- We experiment with different machine learning models, deep neural network models, and fine-tune BanglaBERT, XLM-RoBERTa, and our BanglaClickBERT to develop a Bangla Clickbait Detection model for Bangla textual data. We compare the performance of our model using different metrics.

II. Literature Review

The roots of clickbait can be found in tabloids, a form of journalism that has existed since the 1980s [7]. 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. [8]. Potthast et al. (2016) [9] and Biyani et al. (2016) [10] 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. (2018) [11] mentioned the Clickbait Challenges 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 (2017) [12] 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 (2017) [13] used an LSTM model for the clickbait challenge that included article content. To create the word embedding of clickbait titles, Rony et al. (2017) [14] applied the continuous skip-gram model. Nevertheless, Indurthi et al. (2020) [15] were the first to study the use of transformer regression models in clickbait identification and won the clickbait challenge. Additionally, Hossain et al. (2020) [16] produced the first dataset of Bengali newspapers for Bengali false news detection of around 50K Bangla news articles in an annotated dataset. Apart from Bangla language, clickbaits are used in some different languages news and social media for detection. Genc et al. (2021) [17] used Logistic Regression (85 percent accuracy), Random Forest (86 percent accuracy), LSTM (93 percent accuracy), ANN (93 percent accuracy), Ensemble Classifier (93 percent accuracy), and BiLSTM (97 percent accuracy) on 48,060 headlines from news sources pulled from Twitter for Turkish clickbait detection. Moreover, Balam et al. (2022) [18] used models in the research are Long short-term memory in particular in recurrent neural networks for clickbait detection on social media. Word2vec, use of word embedding vectorize headlines and comparison with Naive Bayes classifier are also performed. Bronakowski et al. (2023) [19] achieved 98 percent 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. We reviewed a paper where a Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN)-based ensemble model is used by Farhan et al. (2023) [20] suggested sarcasm detection AI for Bangla language. achieving 96 percent F1 score and accuracy. We reviewed this paper just to gather information on what type of work can be done using NLP and for gathering some knowledge and examples related to our work. Additionally, for some domain-specific BERT, Beltagy et al. (2019) [21] in order to help enhance efficiency on a range of scientific NLP tasks and produce cutting-edge results, SciBERT which is a pretrained language model based on BERT used unsupervised pretraining on scientific articles. Moreover Jahan et al. (2022) [22] used 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 pre-trained language model in various datasets. However, no research has been done to address clickbait in written news sources using the article's textual properties. Altogether, we will work on BanglaClickBERT which is mainly Bangla clickbait detection from headlines and for that we will use BanglaBERT and MLP techniques.

III. Methodology

We will use Some Statistical Models and Deep Learning Models for setting a benchmark score and then we will implement Transformers Like BanglaBERT, XLM-RoBERTa and Domain Specific BanglaBERT with several variations. Based on this variation we try to come up with the best model.

A. Statistical Models

For statistical methods, 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 (n-gram range=3-5), Bangla pretrained word embeddings, punctuation frequency, and normalized Pars-of-Speech frequency. This will be used as a baseline score.

B. Deep Learning Models

When it comes to deep learning models, there are several powerful techniques that can be employed, such as Convolutional Neural Networks (CNN), 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.

C. BanglaBERT

To tackle the task of clickbait detection in Bangla, we employ the powerful BanglaBERT model [6]. BanglaBERT is a BERT-based Natural Language Understanding (NLU) model pretrained specifically on Bangla using a massive 27.5 GB pretraining corpus.

With its comprehensive understanding of Bangla language nuances, BanglaBERT has demonstrated remarkable performance, surpassing multilingual and monolingual models, and achieving state-of-the-art results across diverse NLP tasks. Given its effectiveness in capturing contextual information and linguistic nuances, we believe BanglaBERT is well-suited for clickbait detection in Bangla.

D. XLM-RoBERTa

We harness the power of XLM-RoBERTa [23] for clickbait detection in Bangla. XLM-RoBERTa, a large-scale multilingual language model based on Facebook's RoBERTa [24].

XLM-RoBERTa undergoes pretraining on an extensive 2.5TB dataset of filtered CommonCrawl data, providing it with a robust and comprehensive understanding of various languages. Unlike certain XLM multilingual models, XLM-RoBERTa does not require language tensors to identify the input language, as it can infer it from the input IDs.

By incorporating RoBERTa tricks within the XLM approach, XLM-RoBERTa focuses on masked language modeling for sentences in a specific language. It achieves impressive performance gains across various cross-lingual transfer tasks, outperforming multilingual BERT (mBERT) [25] on multiple benchmarks.

XLM-RoBERTa excels in handling low-resource languages effectively. It enhances accuracy for languages with limited resources, demonstrating its capabilities on cross-lingual benchmarks such as XNLI.

E. Domain Adaptive Pretraining: BanglaClickBERT

We also propose to further pretrain the model using alarge number of headlines extracted from clickbait-filled websites. Study finds that tailoring pretrained language models to specific domains through adaptive pretraining techniques leads to significant improvements in task performance [26].

This process will involve masked language modeling, a technique where certain words are masked in the input text, and the model learns to predict the masked words based on the surrounding context. Thus, doing so we will convert the BanglaBERT to *BanglaClickBERT*.

With the advent of BanglaClickBERT, this in theory suggest that, our clickbait detection system will witness substantial enhancements, showcasing superior performance in accurately identifying clickbait content from Bangla news headlines.

REFERENCES

[1] S. F. Pengnate, J. Chen, and A. Young, "Effects of Clickbait Headlines on User Responses: An Empirical Investigation," *Journal of International Technology and Information Management*, vol. 30, no. 3, pp. 1-16, 2021. DOI: 10.58729/1941-6679.1440.

[2] Central Intelligence Agency. (2021, June 17). Bangladesh - The World Factbook. *CIA*. https://www.cia.gov/the-world-factbook/countries/bangladesh/

[3] Staff Correspondent and Staff Correspondent, "Prothom Alo at the top with 5 million readers," *Prothomalo*, Mar. 29, 2022. [Online]. Available: https://en.prothomalo.com/bangladesh/prothom-alo-at-the-top-with-5-million-readers

[4] Potthast, M., Gollub, T., Wiegmann, M., Stein, B., Hagen, M., Komlossy, K., Schuster, S., & Fernandez, E. P. G. (2017). Webis Clickbait Corpus 2017 (Webis-Clickbait-17). doi: 10.5281/zenodo.5530410

[5] Mahtab, M., Haque, M., & Hasan, M. (2022). BanglaBait: using transformers, neural networks & statistical classifiers to detect clickbaits in New Bangla Clickbait Dataset. Brac University. http://hdl.handle.net/10361/17128

[6] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, Jul. 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98

[7] S. Elizabeth Bird, *Tabloidization*, *The International Encyclopedia of Communication*, Wiley Online Library, 2008.

[8] M. H. Munna and M. S. Hossen, *Identification of Clickbait in Video Sharing Platforms*, in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pp. 1–6, 2021.

[9] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pp. 810–817, Springer, 2016.

[10] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, "8 amazing secrets for getting more clicks: Detecting clickbaits in news streams using article informality," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[11] M. Potthast, T. Gollub, M. Hagen, and B. Stein, "The clickbait challenge 2017: Towards a regression model for clickbait strength," *arXiv preprint arXiv:1812.10847*, 2018.

[12] Y. Zhou, "Clickbait detection in tweets using self-attentive network," *arXiv preprint arXiv:1710.05364*, 2017.

[13] P. Thomas, "Clickbait identification using neural networks," *arXiv preprint arXiv:1710.08721*, 2017.

[14] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. *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*, pages 232–239, 2017.

[15] V. Indurthi, B. Syed, M. Gupta, and V. Varma, *Predicting clickbait strength in online social media*, in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4835–4846, 2020.

[16] Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. *Banfakenews: A dataset for detecting fake news in bangla*. arXiv preprint arXiv:2004.08789, 2020.

[17] S. Genç and E. Surer, "ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms," *Journal of Information Science*, vol. 49, no. 2, pp. 480–499, 2023.

[18] A. A. Balan, P. Anoop, and A. S. Mahesh, "Clickbait Detection Using Long short-term memory," in *2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS)*, pp. 159–163, 2022, IEEE.

[19] M. Bronakowski, M. Al-khassaweneh, and A. Al Bataineh, "Automatic Detection of Clickbait Headlines Using Semantic Analysis and Machine Learning Techniques," *Applied Sciences*, vol. 13, no. 4, p. 2456, 2023.

[20] N. Farhan, I. T. Awishi, M. H. K. Mehedi, MD. M. Alam, and A. A. Rasel, "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)*, pp. 0624–0629, 2023.

[21] Iz Beltagy, Kyle Lo, and Arman Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[22] Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah, *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, pp. 8–15, 2022.

[23] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 1, pp. 1–1, Jan. 2022, doi: 10.48550/arXiv.1907.11692.

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 1-1, Jan. 2022, doi: 10.48550/arXiv.1810.04805.

[26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 1-1, Jan. 2022, doi: 10.48550/arXiv.2004.10964.