

Bangla Online Clickbait Detection: Constructing Dataset and Use of Transformer Models and Deep learning models

Sifat Ahmed

Department of CSE

Ahsanullah University of Science Technology

Dhaka, Bangladesh

190204006@aust.edu

Aurnab Barua

Department of CSE

Ahsanullah University of Science Technology

Dhaka, Bangladesh

190204009@aust.edu

Mrinmoy Saha Joy

Department of CSE

Ahsanullah University of Science Technology

Dhaka, Bangladesh

190204017@aust.edu

Anika Waziha

Department of CSE

Ahsanullah University of Science Technology

Dhaka, Bangladesh

190204020@aust.edu

Abstract—Clickbait has become a serious issue affecting the reliability of online news and content in languages with limited digital resources, including Bengali. This study aims to address this problem by developing machine learning models for Bangla clickbait detection. We contributed in extending dataset and worked on a dataset of overall 15,000 Bangla news articles, social media posts and Youtube video titles labeled as either clickbait or non-clickbait. Various natural language processing techniques were employed to preprocess the text data, including tokenization, stopwords removal and word encoding. We then trained deep learning sequence classification models, specifically Bidirectional LSTM and transformer-based models like BanglaBERT and Multilingual BERT, on this dataset. Experimental results demonstrate that transformer models achieve superior performance compared to traditional neural networks. Specifically, Multilingual BERT attained 88% accuracy, precision, recall and F1 score, indicating its effectiveness at capturing linguistic patterns in Bangla text for clickbait identification. Overall, this research presents promising methods and resources for detecting misleading online content in low-resource languages through transfer learning approaches. Our labeled dataset and models can also serve as a foundation for future work on trust and fact-checking in Bengali cyberspace.

Index Terms—Deep Learning, BiLSTM, Transformer models, BERT, MBERT, BanglaBERT, Bengali, Clickbait.

I. INTRODUCTION

In languages with limited resources like Bengali, the rise of misleading content on social media and online news platforms is a major concern. Clickbaits are mainly attention-grabbing headlines that not only show one thing and deliver other things but it can also result in a huge disappointment. This research aims to tackle this urgent problem by training Natural Language Processing (NLP) models modified to identify clickbait in Bengali text. This will help the online space for Bengali speakers more reliable and trustworthy.

Clickbait is seriously messing with our ability to trust what we read on the internet. It's making everything seem way more exciting or shocking than it actually is. We are focusing on understanding how the Bengali language works, especially in the context of clickbait, and use deep learning and transformer models to detect clickbaits in a better way so people get accurate info online and create a space where trust is given.

We dive into the research work with the use of the database resource we have found and work our way to create our own enriched labeled database of clickbait and non clickbate data. We will classify examples of content as either clickbait or non-clickbait with the help of deep learning models like-BiLSTM and transfer learning models like- BanglaBERT, multi-lingual bert cased. We will take the necessary steps to preprocess the data, select appropriate models, train the system, and evaluate its performance, providing an accurate clickbait detection method in low-resource languages.

II. LITERATURE REVIEW

Mahtab et al. [1] constructed a Bengali clickbait detection dataset containing 15,056 labeled news articles and 65,406 unlabelled news articles extracted from clickbait-dense news sites where articles were labeled by three expert linguists and includes an article's title, body, and other metadata. They finetuned a pre-trained Bengali transformer model in an adversarial fashion using Semi-Supervised Generative Adversarial Networks (SS-GANs). Among the many comparative studies between the performing models the GAN-BanglaBERT outperforms all other models regarding the accuracy of the model is 82.57%. Hossain et al. [2] proposed an annotated dataset

of 50K news that can be used for building automated fake news detection systems for Bengali language. Additionally, an analysis of the dataset has been provided and developed a benchmark system with state of the art NLP techniques to identify Bengali fake news. To collect a set of authentic news, 22 most popular and mainstream trusted news portals in Bangladesh is selected whereas for collecting fake news they included 3 types of news in their dataset. Which are Misleading/False Context, Clickbait and Satire/Parody. Raju et al. [3] used 32,000 articles headlines, 16,000 each classified as clickbait & non-clickbait from various newsportal. The authors pre-processed the data using NLP preprocessing methods. Two NLP techniques TF-IDF & Word2Vec were used & five machine learning classification algorithms: Logistic Regression, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier were used to predict the results. From that the voting classifier showed the best result from both of the techniques. Rasel et al. [4] tried to categorize Bengali fake news. In the working dataset they classified the data in real and fake news and the fake news were also categorized in misinformation, satirical and clickbait. They collected 500 Bengali fake news data via Internet Archive from ChannelDhaka, Earki, and Motikonho and 1299 fake news from different news portals and e-papers such as Motikonho, Bangaliviralnews, Prothombhor and more. It contains 48678 real news from popular Bengali newspapers like Kalerkantho, Prothom Alo, etc. They worked on multiple Machine Learning models, Deep Neural Networks, and Transformer models like BanglaBERT and m-BERT to attain some comparative results. The best performing models are CNN, CNN-LSTM, and BiLSTM, with the accuracy of 95.9%, 95.5%, and 95.3%, respectively. Sarem et al. [5], first labeled the news as click-bait or non-click bait from 72,321 Arabic tweets from twitter for detecting clickbait headlines on social networks in the Arabic language. Six ML classifiers were implemented: Random Forest (RF), Logistic Regression with Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naïve Bayes (NB), and k-Nearest Neighbor (k-NN). From that SVM showed the best result.

III. DATASET

Mahtab et al.[1] created the first database for Detecting Bengali Clickbait News Articles. They chose many Bengali news article sites and gather news titles from those webpages. Also they gathered clickbait titles from various popular news portal of Bangladesh. The dataset contains 15,056 unique news articles with 9,817 non-clickbait and 5,239 clickbait articles. Hossain et al. (2020) published the first dataset of Bengali fake news where there was 176 data categorized as clickbait in terms of Bengali fake news. After collecting and merging these two datasets we felt the necessity of extending the dataset further more thus we focused on clickbait title on Youtube platform. We made a list of some controversial Bengali Youtube channels and collected their posted video titles. We labeled the data as clickbait on basis of few catagories mentioned by Mahtab et al.[1].

After merging our own dataset with previous datasets we got the following number of instances.

TABLE I
LABEL DISTRIBUTION

Class	Number of instance
Clickbait	5319
Non-CLickbait	9819
Total: 15138	

IV. METHODOLOGY

A. Data Preprocessing

The raw Bangla text data was first preprocessed before training machine learning models. This involved:

- Tokenization: We used word tokenizer to split sentences into tokens. This handles characteristics of Bangla NLP.
- Stopword removal: A list of common Bangla stopwords was collected and removed as they don't contribute much to classification.
- Word encoding: The tokenized sequences were encoded into integers using Keras' Tokenizer. This allows the neural network to process textual input.

B. Data Split

The preprocessed data was then split into training and test sets using Scikit-learn's `train_test_split`. We ensured no data leakage by splitting based on samples rather than before preprocessing. 80% of data was allocated to training and remaining 20% to testing.

C. Bidirectional LSTM

A Bidirectional LSTM model was developed for sequence classification:

- Embedding layer: Transformed the integer encoded sequences into dense vectors of size 100.
- Bidirectional LSTM layers: Two stacked BiLSTM layers with 64 and 32 units respectively to capture contextual information in both directions.
- Output layer: Final sigmoid layer outputs the clickbait probability between 0-1.
- Training: Binary cross entropy loss and Adam optimizer were used. Early stopping monitored validation loss to prevent overfitting.
- Evaluation: Model performance was evaluated on test set using accuracy, precision, recall and F1 score after each epoch of training.

D. BanglaBERT

For BanglaBERT, we used the pretrained BanglaBERT base model and added a classification head on top consisting of a Dropout layer with rate 0.2 and a Dense layer with 1 unit and sigmoid activation. The model was trained with a batch size of 16 for 3 epochs.

E. Multilingual cased BERT

For Multilingual cased BERT (MBERT), we used the same architecture as BanglaBERT by adding a classification head on the pretrained MBERT model. We trained it for 3 epochs with a batch size of 16 using binary cross entropy loss and Adam optimizer.

F. Lime Explainable AI

To gain insights into the cse-buet-nlp’s BanglaBERT pre-trained model’s decisions, we performed model explanations using LIME (Local Interpretable Model-agnostic Explanations). LIME is a technique to explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model. For prediction made by BanglaBERT, LIME perturbed the input text by removing words and observed the change in prediction to understand the important features used by the model. This helped identify the keywords and phrases driving the model’s clickbait/non-clickbait classification for different examples.

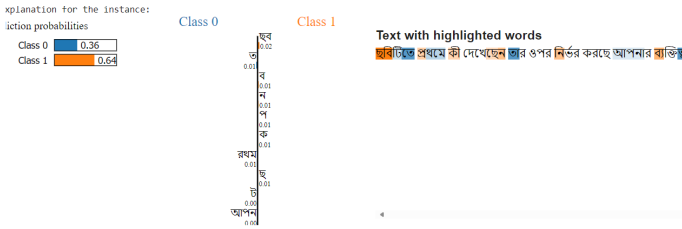


Fig. 1. BanglaBERT model interpretation

V. EXPERIMENTAL RESULT

In this section, we present the experimental results of our study, which focused on identifying clickbait titles on the internet using deep learning and transformer approach. We evaluated the performance of three different approaches: Bidirectional LSTM, BanglaBERT and MBERT. The table below summarizes the results obtained:

TABLE II
RESULT ANALYSIS OF MULTIPLE MODELS

Model	Accuracy	Precision	F1 Score	Recall
Bidirectional LSTM	0.64	0.64	0.50	0.64
BanglaBERT	0.85	0.84	0.85	0.84
MBERT	0.88	0.88	0.88	0.88

VI. RESULT ANALYSIS

As seen from the results in Table II, MBERT achieved the best performance with an accuracy of 88%, precision of 88%, recall of 88% and F1 score of 88% on the test set, followed by BanglaBERT. This could be attributed to the fact that MBERT has been trained on a huge multilingual corpus and has learned better generalizable representations. BanglaBERT being specifically trained for Bengali, performed better than the BiLSTM model. The BiLSTM model achieved an accuracy of only 64% indicating that deep transformer

models are able to capture better linguistic patterns for the task of clickbait detection. Overall, our experiments demonstrate that transformer-based models particularly MBERT can be effectively used to identify clickbaits in Bengali text.

VII. CONCLUSION

To conclude we can state that in our work we tried to understand the Bengali language click-bait structures while classify them and put on a comparison between a deep learning model and pre-trained BERT models. Where mBERT outperform other methods with the highest accuracy of 88%. This work will be beneficial for people in internet as they will be classify and identify the click-bait Bengali contents and browse safe internet.

REFERENCES

- [1] Md. Motahar & Monirul Haque. BanglaBait: Semi-Supervised Adversarial Approach for Clickbait Detection on Bangla Clickbait Dataset
- [2] Hossain, Md & Rahman, Md & Islam, Md Saiful & Kar, Sudipta. (2020). BanFakeNews: A Dataset for Detecting Fake News in Bangla.
- [3] Rajapaksha, Praboda. (2020). Clickbait detection using multimodel fusion and transfer learning.
- [4] Raju, N.V. & Nyalakanti, Nikhil & Kambampati, Premsai & Kanthali, Yeshwanth & Pandey, Shivam & K, Maithili. (2023). Clickbait Post Detection using NLP for Sustainable Content. E3S Web of Conferences. 430. 10.1051/e3sconf/202343001081.
- [5] Saha, Sourav & Sarker, Aditi & Chakraborty, Partha & Yousuf, Mohammad. (2022). Bengali Fake News Detection: Transfer Learning Based Technique with Masked LM Process by BERT. 10.1007/978-3-031-20977-2_7.
- [6] Hossain, Md. Muzakker & Awosaf, Zahin & Prottoy, Md. Salman & Alvy, Abu & Morol, Md. Kishor. (2022). Approaches for Improving the Performance of Fake News Detection in Bangla: Imbalance Handling and Model Stacking.
- [7] Rasel, Risul Islam & Zihad, Anower & Sultana, Nasrin & Hoque, Moshui. (2022). Bangla Fake News Detection using Machine Learning, Deep Learning and Transformer Models. 959-964. 10.1109/IC-CIT57492.2022.10055592.
- [8] Wotaifi, Tahseen & Dhannoon, Ban. (2023). Developed Models Based on Transfer Learning for Improving Fake News Predictions. JUCS - Journal of Universal Computer Science. 29. 491-507. 10.3897/jucs.94081.
- [9] Al-Sarem, Mohammed & Saeed, Faisal & Al-Mekhlafi, Zeyad & Hadwan, Mohammed & Al-Hadhrani, Tawfik & Alshammari, Mohammad & Alreshidi, Abdulrahman & Alshammari, Talal & Al-Shaibani, Badiea. (2021). An Improved Multiple Features and Machine Learning-Based Approach for Detecting Clickbait News on Social Networks. Applied Sciences. 11. 9487. 10.3390/app11209487.
- [10] "Bhattacharjee, Abhik and Hasan, Tahmid and Ahmad, Wasi and Mubasshir, Kazi Samin and Islam, Md Saiful and Iqbal, Anindya and Rahman, M. Sohel and Shahriyar, Rifat". BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla".
- [11] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.