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

Team No: 32

Members:
Saman Sarker Joy - 20101114
Tanusree Das Aishi - 20101012
Naima Tahsin Nodi - 20101150

ST: Sadiul Arefin Rafi
RA: Md Sabbir Hossain

Introduction

- The rise of online news media and clickbait titles
- Lack of clickbait detection research in Bangla language
- Challenges and importance of clickbait detection

খালের মাঝে টানা জাল টান দিতেই ওঠলো বিশাল বড় চিতই মাছ, জেলেদের মাছ ধরার ভিডিও ভাইরাল

As the nets drag in the canal, the big chitfish, the fishing video went viral

আমার স্ত্রী প্রাইমারি স্কুলের টিচার, একদিন রাতে ডিনারের শেষে ...

My wife is a primary school teacher, one night at dinner ...

রেমান্ডে যাদের বিষয়ে গুরুত্বপূর্ণ তথ্য দিলেন পরীমনি

Porimoni gave important information about those on remand

Literature Review

- Evolution and history of clickbait
- Existing research in clickbait detection in other languages, especially in English
- Limited research in Bangla clickbait detection
- Use of different techniques and models in related tasks

Problem Statement

- Formulating clickbait detection as a binary classification task with two main categories

  $C = \{clickbait,\ non - clickbait\}$
- The aim of the research: Detecting clickbait in Bangla news headlines
- The need for accurate and efficient clickbait detection models

So, the problem can be formulated as,

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

Dataset Description

Annotated Dataset

- 15,056 labeled Bangla news articles
- Categorized as clickbait (1) or non-clickbait (0)
- Collection period: Feb 2019 - Feb 2022

Unannotated Dataset

- 65,406 unannotated Bangla news articles from clickbait-dense websites
- Expanded it to around 1 million headlines using by scraping clickbait dense news portals
- Utilized for pretraining BanglaBERT model

Methodology

Statistical Models

- Logistic and Random Forest classifiers
- Features: TF-IDF, word embeddings, punctuation frequency, POS frequency

Deep Learning Models

- BiLSTM, Ensemble Models
- Effective for text classification tasks

Methodology Cont.

Transformer Models

❖ **BanglaBERT**

- Pretrained on 27.5GB Bangla corpus
- State-of-the-art results in various NLP tasks

❖ **XLM-RoBERTa**

- Multilingual model based on RoBERTa
- Pretrained on extensive multilingual data

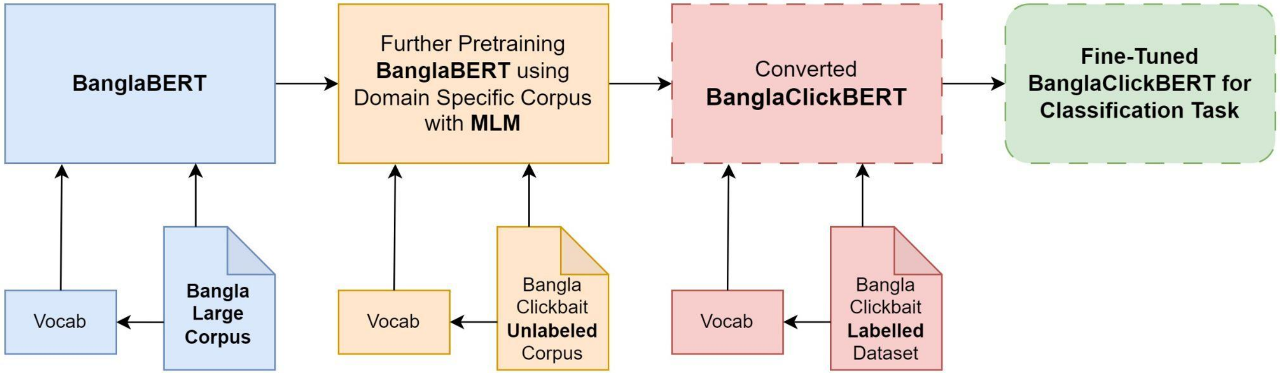
❖ **Domain Adaptive Pretraining: BanglaClickBERT**

- Further pretraining BanglaBERT using headlines from clickbait-filled websites
- Adaptive pretraining for domain-specific improvements

Creation of BanglaClickBERT

- **Reason for Pretraining**
  - Don't Stop Pretraining Paper
  - BanglaHateBERT Paper
- **Pretraining Data**
  - Unannotated dataset
- **Training Strategy**
  - MLM (Masked Language Modelling)
  - Epoches 10, LR 5e-5, Seq Length 32
  - Took us almost 28 hours to pretrain

Creation of BanglaClickBERT Cont.



Model Available: https://huggingface.co/samanjoy2/banglaclickbert_base

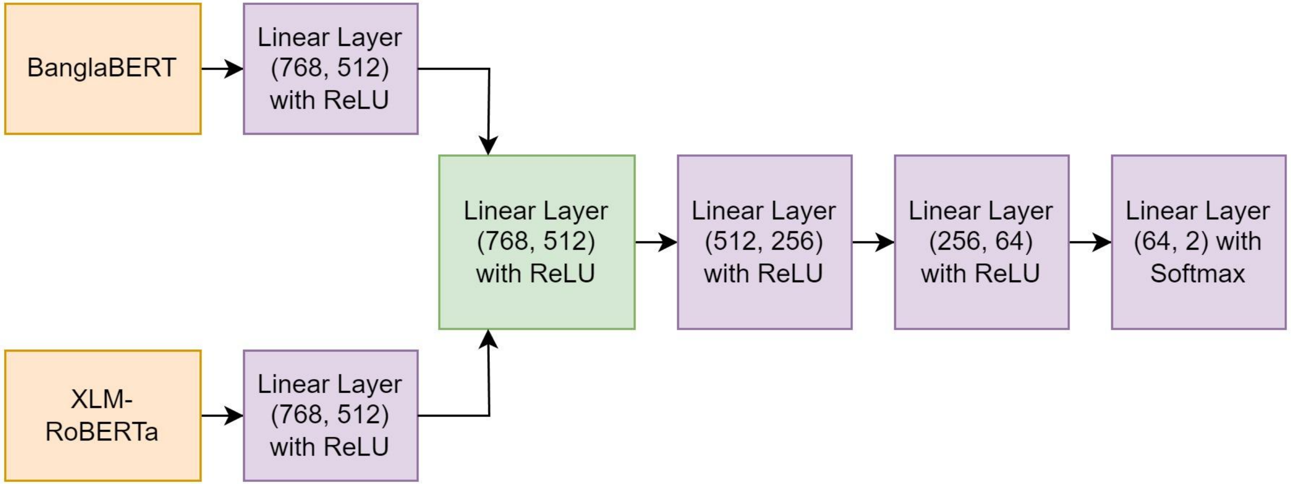
System Overview

- **Statistical Models :** Logistic Regression, Random Forest
- Used TF-IDF vectors to capture the sequential patterns of characters in the text
- **Deep Learning Models :** Bi-LSTM Network model, Ensemble of Convolutional neural network + Gated recurrent unit
- Both implemented with Bengali GloVe Embeddings

System Overview Cont.

- Experimented with various **Transformer model Configurations**
    - **BanglaBERT / XLM-RoBERTa / BanglaClickBERT (last layer) + MLP:** the last layer of the BanglaBERT and XLM-RoBERTa base models is used as the input
    - **BanglaBERT / XLM-RoBERTa / BanglaClickBERT (average of all layers) + MLP:** takes the average of all layers in the BanglaBERT and XLM-RoBERTa base models instead of only the last layer
    - **BanglaBERT / BanglaClickBERT and XLM-RoBERTa concatenation of the last layer + MLP**: the outputs from the last layers of BanglaBERT and XLM-RoBERTa are concatenated together

System Overview Cont.



Experimental Setup

- Prepossessing
  - Dataset was already preprocessed
  - We used *bnunicodenormalizer*
    - Removed any broken unicodes
    - Removed any English Texts
- Experimental Settings
  - N-grams length from 1 to 5
  - 300d Bangla GloVe embeddings
  - Used all base models for

    Transformers (12 layers)

Experimental Setup Cont.

- Hyperparameters
  - 20 Epoches, LR 1e-5, Maximum length 32, Batch size 128
  - Cross Entropy Loss
  - Optimizer AdamW
- Dataset Splits
  - 70% (10839 headlines) training
  - 20% (3012 headlines) testing
  - 10% (1205 headlines) validation
- Metrics
  - Precision, Recall, macro F1-Score and Accuracy

Results and Analysis

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	0.7513	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*. The models that used BanglaClickBERT have shown consistent results than other models.

Results and Analysis Cont.

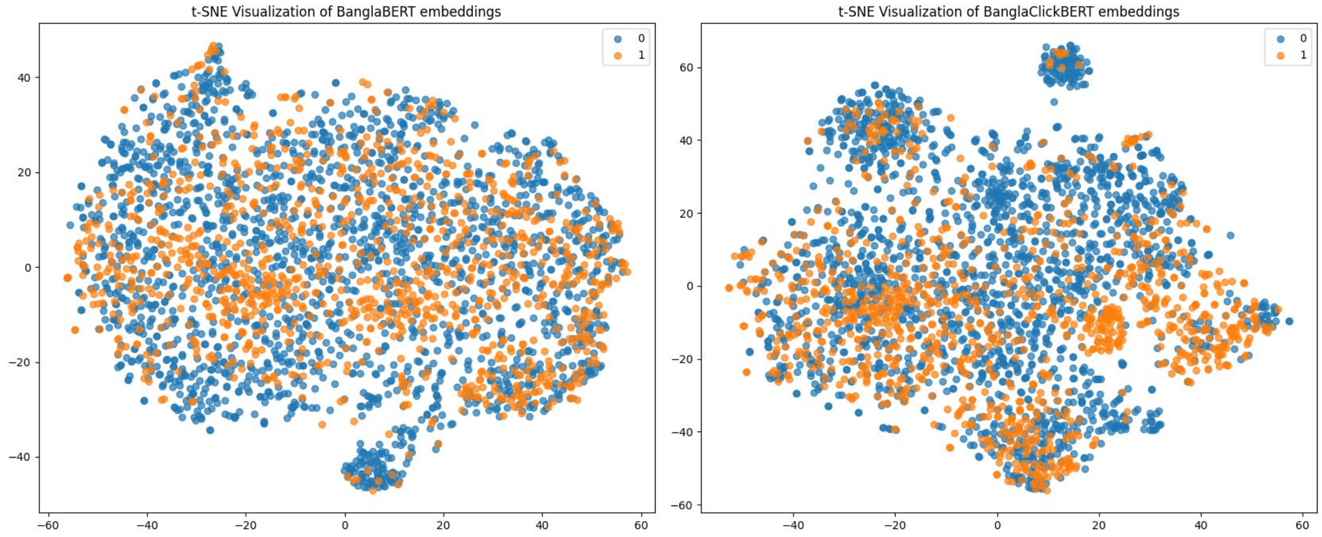


Figure: Visualization of last layer hidden representations using t-SNE (van der Maaten and Hinton, 2008) for BanglaBERT (Left) and BanglaClickBERT (Right) without any fine-tuning. 0 represents Non-Clickbait and 1 represents Clickbait in both figures

Conclusion

- A significant advancement in the field of clickbait detection in Bangla
- Augmented the unlabelled dataset
- Created a sophisticated solution for this problem
- Created Domain Adaptive BanglaClickBERT and made it publicly available

Future Works

- Use Large Transformer Models
- Make more labelled dataset in this domain

References

- Jahan, M. S., Haque, M., Arhab, N., & Oussalah, M. (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 (pp. 8-15). Marseille, France: European Language Resources Association.
- Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., & Shahriyar, R. (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 (pp. 1318-1327). Seattle, United States: Association for Computational Linguistics. doi:10.18653/v1/2022.findings-naacl.98
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (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 (pp. 8342-8360). Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.740
- And more…..

Thank You.