From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages with LLMs

Anonymous submission

Abstract

Fake news is a persistent global challenge in an era of rapid information dissemination, demanding swift and effective intervention. While manual fact-checking remains the gold standard, its time-consuming nature leaves ample room for misleading content to inflict substantial damage. This predicament is intensified for low-resource languages like Bangla, which need more datasets and tools for efficient fake news detection. In this paper, we proposed BanFakeNews-2.0, generated from various newspaper and online platforms with 13 categories of 47k authentic news and 13k meticulously curated and manually annotated fake news in Bangla. Notably, BanFakeNews-2.0 is an order of magnitude larger than existing fake news datasets. To understand the data characteristics, we conduct exploratory analysis of BanFakeNews-2.0 and develop a benchmark system with state-of-the-art NLP techniques to identify Bangla fake news. Moreover, we provide a benchmark with a transformer-based model BERT, its variants in Bangla, and multilingual. We fine-tuned BLOOM using the QLORA, using gradient accumulation from each step and a paged Adam 8-bit optimizer for the classification task. We demonstrated that the large language model(LLM)–based and transformer-based models outperform the traditional linguistic features and neural network-based methods. BanFakeNews-2.0 is shown to have great potential to prompt fake news studies in Bangla.

Keywords: Fake News, Fact Check, Bangla, NLP

1. Introduction

The proliferation of fake news, defined as the intentional dissemination of false information, has emerged as a pervasive concern in modern society. Such misleading narratives, propagating through diverse channels such as social media, online news platforms, and even traditional newspapers, serve not merely as innocent fallacies but as instruments to deceive, influence, and sometimes manipulate public perception. The consequences of such disinformation can range from shaping public opinion on critical matters to catalyzing large-scale societal disturbances (Rapp and Salovich, 2018; Grinberg et al., 2019; Pandey, 2019). An emblematic instance of its potential repercussions surfaced in 2020 when misinformation regarding the COVID-19 vaccine's adverse effects gained traction on social media, leading to widespread vaccine hesitancy and skepticism (Lee et al., 2022; O'Connor and Murphy, 2020). A particularly poignant study from Bangladesh (Shirina and Prodhan, 2020) illuminated the dangerous consequences of misinformation in digital domains, enumerating various incidents from baseless child kidnapping rumors to unwarranted religious hate speech. The 2012 episode involving unwarranted attacks on Buddhist temples in Bangladesh, instigated by groundless accusations on a social media platform, underscores the scale of turmoil such falsehoods can unleash (Bhikkhu, 2014). The manifestations of fake news are multiform, spanning textual articles, images, videos, and even memes.

The contemporary strategy to solve this misinformation challenge primarily relies on manual verification by specialized entities, a method that necessitates enhanced scalability. While there's an increasing inclination towards harnessing automated computational methodologies for this purpose, it's evident that the preponderance of research efforts is oriented towards the English language. Furthermore, the requirement for manually annotated fake news corpora remains a substantial impediment, hindering the progression and implementation of computationally sophisticated, wide-coverage models in this domain. Hossain et al. (2020) pioneered in this realm by introducing a public dataset, 'BanFakeNews,' albeit it encompasses merely 100 instances of fake news, supplemented with 900 satirical news pieces. Subsequently, Al-Zaman and Noman (2023) unveiled a more expansive dataset comprising 4.5k fake news items spanning both Bangla and English. Meanwhile, Sharma et al. (2019) achieved a 96% accuracy rate with a hybrid technique for satirical news detection. Sraboni et al. (2021) evaluated various classifiers, spotlighting the potential of the aggressive classifier and SVM. Moreover, Hussain et al. (2020a) found the SVM with a linear kernel outperformed MNB in Bangla fake news detection. Our findings indicate that Bangla fake news research is still nascent, prompting us to develop a new dataset to bolster further studies.

Nevertheless, conspicuously absent from the current research landscape are endeavors employing state-of-the-art Large Language Models (LLMs) or BERT-centric methodologies, particularly for under-resourced languages. This manifests as a pronounced gap in the literature, especially given the pressing need for comprehensive resources and frameworks tailored for languages like Bangla to mitigate the spread of fake news effectively. In

Dataset source	#FN	#TN
SadikAlJarif (2022)	4.5K	10K
Al-Zaman and Noman (2023)	2K	5k
Hossain et al. (2020)	1.3K	48.6k
Hussain et al. (2020b)	1K	2.5K
Our	13K	47k

Table 1: Existing fake news dataset in Bangla, here #FN represents no of Fakenews and #TN represents the no of true news dataset

addition, LLM and transformer-based model performance is missing in existing literature. To address these issues, we introduce BanFakeNews-2.0, a significantly upgraded version of BanFakeNews (Hossain et al., 2020), where we follow similar data collection, annotation, and validation approaches. However, collect more data from various categories - medical, religious, and fake news. We have made available a publicly accessible dataset of nearly 60K annotated Bangla news articles, including 13k fake news. With such volume and a period of a decade, LIAR is an order of magnitude larger than the currently available resources (Hossain et al., 2020; SadikAlJarif, 2022; Al-Zaman and Noman, 2023; Hussain et al., 2020b) of similar type, details are shown in Table 1.

Empirically, we have evaluated several state-of-the-art learning-based methods, including traditional linguistic features and neural network-based methods, transformer-based models, and LLMs. Our experiment suggests that the based approach improves performance compared to baseline methods, including linguistic features and neural network-based methods. We anticipate that this research will be instrumental in advancing fake news detection systems. Throughout the remainder of the paper, we provide a concise overview of our dataset preparation procedures and experiment with baseline and transformer-based models and findings.

2. Dataset

We are introducing a newly created and annotated Bangla Fake News dataset consisting of nearly 13000 fake(labeled as 0) and 47000 authentic news(labeled as 1) and content collected from different online news portals and mainstream news media. To collect authentic news articles, we selected the top 30 news portals in Bangladesh, renowned for their credibility and widespread readership. Simultaneously, to curate fake news, we identified six predominant fact-checking platforms extensively employed for debunking misinformation in the Bangladeshi context. Recognizing the variegated web structures of these platforms, we designed and implemented an automated web crawler tailored for each fact-checker. This ensured comprehensive data extraction across diverse webpage layouts, which was subsequently consolidated into a singular database. Furthermore, our dataset encapsulates various shades of fake news and misinformation, encompassing misleading or false context, clickbait, and satirical news genres, thus providing a holistic view of the misinformation landscape. Post-collection, we undertook rigorous preprocessing to enhance the data quality, eliminating NaN values and excising duplicate news entries. A significant portion of our fake news was sourced from jachai[1] and boombd[2]. Both platforms offer erudite elucidations of the counterfeit news that has pervasively permeated the Bengal region via diverse online media. Specialized web scraping tools and Python scripts propelled our manual data aggregation. We scrutinized many potentially misleading websites, with the aforementioned two platforms being primary contributors. This manual curation entailed:

- Pinpointing pages replete with counterfeit news.
- Excluding antiquated news pages (already accounted for in our other collections).
- Ascertain the HTML elements housing the pertinent news headline and content.
- Fine-tuning the web scraper for analogous pages and subsequently executing the data extraction.

As a foundation for the data collection, annotation, and compilation, we predominantly utilized the benchmark dataset BanFakeNews Hossain et al. (2020) process, culminating in an aggregate of approximately 47,000 genuine news articles. In the data collection and annotating process, the authenticity of the source and credibility of both real and fake news has been assured. The Faculty of Applied Science and Engineering undergraduate students have carefully conducted this process.

2.1. Dataset Statistics

In our dataset analysis, we observed a myriad of categories, attributable to the distinct classification methodologies adopted by various news publishers and distributors of misinformation. For standardization, we consolidated analogous types across different news sources into unified categories. Consequently, all news items have been grouped into 13 distinct categories. Our aim was to accrue up to 500 fake news articles for each category to ensure a balanced dataset. However, attaining this

[1]www.jachai.org
[2]www.boombd.com

count proved challenging for certain categories, notably lifestyle, medical, and religious, though we endeavored to include as many articles as feasible for these sections. In total, the dataset comprises 60,000 news articles. The distribution across each category is delineated in Table 2.

Category	Authentic	Fake
Politics	2941	3022
Miscellaneous	2218	1556
International	6990	1395
Lifestyle	901	241
Medical	0	448
Religious	0	300
Sports	6526	884
Educational	1115	787
Technology	843	688
National	18708	1140
Crime	1072	551
Entertainment	2636	1405
Finance	1224	544

Table 2: Number of news in each news category

3. Methodologies

In this section, we elaborate on our devised systems for the identification of fake news in Bangla. Our methodologies encompass both conventional linguistic attributes and employ a range of models based on neural networks.

Traditional Approaches: We extracted lexical linguistic features utilizing the TF-IDF of the character n-grams (n = 3,4,5) and word n-grams (n = 1,2,3). Linear Support Vector Machine (SVM) (Hearst et al., 1998) was used on the features to get outputs.

Transformer-based BERT models: Encoder-based pre-trained BERT (Devlin et al., 2018a) models are exceptional in downstream tasks due to their superior contextual understanding capabilities. We chose five pre-trained model bases: BanglaBERT (Bhattacharjee et al., 2022) and Bangla-BERT by Sarker (2020) (later on referred to as Sagor-BERT), which are monolingual, XLM-RoBERTa (late referred as XRoBERTa) (Conneau et al., 2019), multilingual-BERT cased and uncased (later on referred as m-BERT-c and m-BERT-unc respectively) by Devlin et al. (2018b) which are multilingual. We shuffled the training samples and enforced gradient clipping to fine-tune these models. We utilized the outputs from the last two layers of multi-head attention, subsequently employing a linear layer for classification. We fine-tuned the model using Adam optimizer (Kingma and Ba, 2014).

Large Language Model BLOOM: Large language models (LLM) have recently demonstrated remarkable proficiency in the realm of linguistic analysis and reasoning. BLOOM (Scao et al., 2022) is a decoder-based LLM, a product of the most extensive single project collaboration of AI researchers. We used the BLOOM 560 million parameters model with a linear classification head on top, loaded and fine-tuned using the QLORA (Dettmers et al., 2023) for the classification task. We used gradient accumulation from each step and a paged Adam 8-bit optimizer for fine-tuning.

4. Experimental Setup

Data Pre-processing: English words and hyperlinks were removed from the dataset. Text normalization, punctuation, and stop-words removal were performed for traditional models. As punctuation is essential for capturing context in a sentence, there was no punctuation removal for our experimentation with BERT and BLOOM.

Model Validation and Dataset Split: We validated the models using the holdout method. For this purpose, we split the dataset into train, validation, and test sets containing 70%, 15%, and 15%, respectively, while keeping the same class ratio.

Baselines: In our experimental evaluation, we benchmark our results against two baseline approaches. Firstly, a majority baseline assigns the predominant class label (in this case, 'authentic news') to all articles. The second is a random baseline, which randomly classifies articles as authentic or fake. Table 3 presents the average precision, recall, and F1-score obtained from 10 random baseline experiments.

Experiments: For each experiment, we chose the hyperparameters based on the validation set and evaluated the model on the test set. For traditional models, we only trained on the content of the news.

For BERT and BLOOM, we trained both on content and headline while keeping a maximum of 512 input tokens limit. To differentiate the headline and content of each news sample, we added the string " \\ "between these.

5. Result and Analysis

All the experiment results (in percentages) are shown in Table 3. In our approach, performances were validated using the holdout method leading more unbiased performance measure than the previous similar works in Bangla. The Precision(P), Recall(R), F1(F1-Score) of the authentic class, and the P of the fake class are quite high, while the R of the authentic class is almost perfect. In most cases, we achieve more than 90% Precision, Recall, and F1 for authentic class. However, the results of Precision, Recall, and F1-Score of fake

class vary from experiment to experiment. While the experiment with linguistic features with SVM outperforms LLM and transformer-based models in classifying authentic news, LLM-based models outperform SVM models in classifying fake news with higher F1 scores. Transformer models (m-BERT-uncased (m-BERT-unc) and BLOOM) performed quite well in detecting fake with an F1 score of 81% compared to the fake class F1 score of 77% by C3-Gram. On the other hand, the transformer models, highest F1 being 96, performed slightly worse than traditional models whose highest F1 score was 98. This can be credited to the significant increase of fake news in the dataset compared to all others. We see that LLM BLOOM and m-BERT-uncas performed the best in all aspects compared to other transformer models. SagorBERT, m-BERT-c(cased), m-BERT-unc(uncased), and BLOOM performed similarly in all aspects except for P of the fake class. Conversely, BanglaBERT fell behind due to its low P and R for authentic and fake classes respectively.

Model	**Authentic**			**Fake**		
	P	**R**	**F1**	**P**	**R**	**F1**
Baselines						
Majority	79	100	88	0	0	0
Random	79	50	61	21	51	30
Linguistic Features with SVM						
Unigram(U)	92	95	93	78	70	74
Bigram(B)	91	95	93	78	67	72
Trigram(T)	91	88	90	62	69	66
U+B+T	92	95	94	79	70	75
C3-Gram(C3)	96	97	98	80	74	77
C4-Gram(C4)	97	98	97	79	75	77
C5-Gram(C5)	96	97	96	81	74	77
C3+C4+C5	97	98	97	79	75	77
BERT models						
BanglaBERT	89	99	94	97	53	69
SagorBERT	92	99	95	95	68	79
m-BERT-c	92	98	95	93	69	79
m-BERT-unc	92	99	96	99	69	81
XRoBERTa	90	98	94	89	61	72
LLM						
BLOOM	92	100	96	99	69	81

Table 3: Precision (P), Recall (R), and F1 score for each categorical class (Authentic and Fake)

Among linguistic features, the C3-Gram model performed the best among character-based linguistic feature models and unigram+bigram+trigram (U+B+T) performed best among word-based feature models. Character-based linguistic features outperformed word-based features in fake news detection scoring higher performance where C3-Gram performed 1%, 4%, and 2% higher P, R, and F1 respectively compared to U+B+T features. C3-gram also outperformed U+B+T in authentic news identification, indicating higher performance of character-based linguistic features compared to word-based linguistic features in fake news detection.

6. Conclusion

In this study, we introduced the most comprehensive and robust Bangla fake news dataset with 13k manually annotated fakenews. Relative to earlier datasets, it encompasses nearly every vital news category we encounter daily, enhancing its size in real-world relevance and applicability in research. Our evaluation encompassed traditional linguistics feature-based, BERT-based, and LLMs-based models. We fine-tune BLOOM with QLORA, employing gradient accumulation at every step and leveraging a paged Adam 8-bit optimizer for the downstream task in Bangla. It highlights that the Large Language Model (BLOOM) and BERT model m-BERT-unc notably outpaced their counterparts in performance. Interestingly, while evaluations on the BanFakeNews dataset by Hossain et al. (2020) favored traditional models, introducing a more diverse array of fake news saw transformer models gaining an edge. This underscores the significance of context, especially as datasets become more diverse, overshadowing character-based linguistic features. A notable observation was the prevalence of punctuation marks, escape characters, and HTML tags within our fake news dataset—challenges adeptly addressed by our chosen evaluation models. The persistent spread of fake news mandates ongoing surveillance and mitigation, emphasizing the need for more balanced and diverse datasets. Future directions for the dataset presented herein encompass refining features and models, crafting real-time monitoring mechanisms, enhancing annotations, and broadening the dataset. Additionally, exploring zero-shot classification using burgeoning LLMs like LLAMA 2 (Touvron et al., 2023) and GPT-3 (Brown et al., 2020) could offer compelling insights into their efficacy in detecting fake news. In future, refining the dataset's features, enhancing model capabilities, and exploring contemporary Large Language Models could pave the way for more robust fake news detection mechanisms. We remain optimistic that our contributions will bolster further research endeavors in the domain of Bangla fake news detection and mitigation.

7. Bibliographical References

Adnan Ahmad and Mohammad Ruhul Amin. 2016. Bengali word embeddings and it's application in solving document classification problem. In *2016*

19th international conference on computer and information technology (ICCIT), pages 425–430. IEEE.

Md. Sayeed Al-Zaman and Mridha Md. Shiblee Noman. 2023. A dataset on social media users' engagement with religious misinformation. *Data in Brief*, 49:109439.

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.

Pragyananda Bhikkhu. 2014. Who will be tried for ramu destruction? Published: 30 Sep 2014, 16: 58.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Olga Filatova. 2016. More than a word cloud. *Tesol Journal*, 7(2):438–448.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

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.

Md Gulzar Hussain, Md Rashidul Hasan, Mahmuda Rahman, Joy Protim, and Sakib Al Hasan. 2020a. Detection of bangla fake news using mnb and svm classifier. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 81–85. IEEE.

Md Gulzar Hussain, Md Rashidul Hasan, Mahmuda Rahman, Joy Protim, and Sakib Al Hasan. 2020b. Detection of bangla fake news using mnb and svm classifier. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 81–85.

Md Aktarul Islam, Md Sajjat Hossain, Md Tabiur Rahman Prodhan, and Md Abu Bakar Siddique. 2023. Spreading fake content via social media among tertiary level students in rangpur, bangladesh.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, Sun Kyong and Sun, Juhyung and Jang, Seulki and Connelly, Shane. 2022. *Misinformation of COVID-19 vaccines and vaccine hesitancy*. Nature Publishing Group UK London.

P McCullagh and J Nelder. 1989. Generalized linear models.,(chapman & hall/crc: London.).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Cathal O'Connor and Michelle Murphy. 2020. Going viral: doctors must tackle fake news in the covid-19 pandemic. *Bmj*, 369(10.1136).

Geeta Pandey. 2019. The deadly germs lurking in your makeup bag. *BBC News*. Accessed: [20-10-2023].

David N Rapp and Nikita A Salovich. 2018. Can't we just disregard fake news? the consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, 5(2):232–239.

SadikAlJarif. 2022. bangla fake news dataset. https://www.kaggle.com/datasets/sadikaljarif/bangla-fake-news-detection-dataset?select=final_bn_data.csv.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Arnab Sen Sharma, Maruf Ahmed Mridul, and Md Saiful Islam. 2019. Automatic detection of satire in bangla documents: A cnn approach based on hybrid feature extraction model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE.

Sharifa Umma Shirina and Md. Tabiur Rahman Prodhan. 2020. Spreading fake news in the virtual realm in bangladesh: Assessment of impact. *Global Journal of Human-Social Science*, 20(A17):11–25.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Tasnuba Sraboni, Md Rifat Uddin, Fahim Shahriar, Ruhit Ahmed Rizon, and Shakib Ibna Shameem Polock. 2021. *FakeDetect: Bangla fake news detection model based on different machine learning classifiers*. Ph.D. thesis, Brac University.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.