A Corpus and Machine Learning Models for Fake News Classification in Bengali

Yudhanjaya Wijeratne, Masudur Rahman, Kamal Hosen, Munir Hossain, Shahnoor Wahid
LIRNEasia, 12 Balcombe Place, Colombo, Sri Lanka (yudhanjaya@lirneasia.net)

LIRNEasia is a pro-poor, pro-market think tank whose mission is *catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Abstract

We present a dataset consisting of 3468 documents in Bengali, drawn from Bangladeshi news websites and factchecking operations, annotated as CREDIBLE, FALSE, PARTIAL or UNCERTAIN. The dataset has markers for the content of the document, the classification, the web domain from which each document was retrieved, and the date on which the document was published. We also present the results of misinformation classification models built for the Bengali language, as well as comparisons to prior work in English and Sinhala.

Introduction

The ability to computationally detect and classify misinformation at scale is useful in today's world. As posited by prior work Wijeratne and Attanayake [1] and Jayawickrama et al. [2], natural language processing and machine learning techniques exist that show a high degree of accuracy at performing this kind of text classification, even accounting for differences in language structures. Nervertheless, lack of data hamstrings attempts to built tooling for such tasks, especially in languages other than English (Duong, 2017 [3], Wijeratne et al, 2019[4], Wijeratne et al, 2021[1]). This issue affects the Bengali language as well.

Thus we present an annotated dataset of Bangladeshi text for misinformation classification, and the results of machine learning models, conducted in a manner similar to Jayawickrama et al. [2].

Methodology

Data acquisition and cleaning

A primary dataset of 3469 news articles were scraped from Bangladeshi news websites, both reputed and of lesser renown. To maintain parity with prior work in Sinhala([2], articles consisting exclusively of non-textual media were dropped in the process of scraping itself. This we consider the primary dataset. Based on previous findings from the same, a further 510 rumors were manually gathered from the websites of factcheckers operating in Bangladesh. This we consider the secondary dataset. Both datasets, put together, originate from 180 sources, spanning social media to news websites to blogs.

Each data point in this dataset consisted of three variables.

1. domain: The source from which the article was obtained
2. datestamp: The date on which the article was published on the source
3. content: Textual content of the news article

Data annotation

The sample of <x> articles from the primary dataset was divided equally among three annotators, each of whom are trained journalists working in Bangladesh. Each news article was categorized as CREDIBLE, FALSE, PARTIAL, or UNCERTAIN, as per this annotation schema.

1. CREDIBLE : Articles that could be verified to be completely credible
2. FALSE : Articles of which the main argument could be verified to be false
3. PARTIAL : Articles of which the main argument could be verified to be credible, but one or more minor facts could be verified to be false
4. UNCERTAIN : Articles that could not be verified to belong to any of the previous categories

Prior work in Sinhala ([2] incorporated local expertise in the form of a tier list built out of a factchecker's working expertise. For this dataset, we chose to rely on the journalist's judgement as well as that of a senior editor of considerable years of experience in the field, who oversaw the process and quality.

For an article to be annotated as CREDIBLE, FALSE or PARTIAL, a minimum of two sources were required to have published information that could be used to verify or disprove the contents of the article. If sufficient evidence could not be gathered to classify an article as CREDIBLE, FALSE, or PARTIAL, it would be classified as UNCERTAIN.

To ensure better representation of data classes, the secondary dataset gathered from factcheckers was added to supplement the FALSE category. Duplicates were removed, as were corrupt data. With this, the resultant annotated dataset comes to a total of 3468 documents: 2157 CREDIBLE, 774 UNCERTAIN, 517 FALSE, and 20 PARTIAL.

Machine learning for task fitness

Wijeratne et al (2021)[5] demonstrated that classical machine learning, in English, can readily achieve >90% accuracy in misinformation classification. Subsequent experiments in Sinhala [2] showed that similar results, using the same algorithms, could be readily achieved in Sinhala, albeit not to as high a degree as the English results. The Sinhala experiment also suggested that 'modelling uncertainty' - ie: classifying between CREDIBLE and UNCERTAIN instead of the classical true-false dichotomy embodied in CREDIBLE and FALSE - might also be a viable method for a practical implementation.

Accordingly, mimicking the Sinhala test as close as possible, the Bengali dataset presented here was partitioned into a number of class-balanced datasets:

1) CREDIBLE-FALSE 2) CREDIBLE-UNCERTAIN 2) UNCERTAIN-FALSE

517 documents for each class were used. All data was cleaned of punctuation and non-Bengali characters. The algorithms mentioned in our previous work - Naive Bayes, Logistic Regression, Random Forests, eXtreme Gradient Boosting, and Support Vector Machines - were run using

the same 80-20 training-testing data split, with each algorithm run being repeated five times using random sampling of the dataset for training data. The results are the averages of five runs.

Model Used	**F1 Score**	
	Credible	**False**
Naive Bayes	0.73	0.33
Logistic Regression	0.90	0.87
Random Forests	**0.94**	**0.93**
eXtreme Gradient Boosting	0.92	0.91
Support Vector Machines	0.92	0.91

Table 1: Average results of binary classification, CREDIBLE vs FALSE

Prior work in Sinhala contains similarities to our observations with the Bengali data (notably, the representation of overtly FALSE news from the primary dataset is low). However, it would appear that classification accuracy, using the secondary dataset to supplment, is more accurate here, rivalling or surpassing comparable benchmarks in English using the same algorithms[1]. Whether this is due to differences between language structures, or due to the non-removal of stopwords from the Bengali text, or an artefact of this particular dataset, is difficult to establish without further research.

Model Used	**F1 Score**	
	Credible	**Uncertain**
Naive Bayes	0.76	0.74
Logistic Regression	0.75	0.74
Random Forests	0.75	0.75
eXtreme Gradient Boosting	0.74	0.75
Support Vector Machines	**0.76**	**0.78**

Table 2: Average results of binary classification, CREDIBLE vs UNCERTAIN

Prior research shows that the accuracy of such models drop when classifying more classes. For a multi-class classification, 568 articles from the CREDIBLE, FALSE and UNCERTAIN categories were used and the above steps performed again.

Model Used	**F1 Score**		
	Credible	**False**	**Uncertain**
Naive Bayes	0.57	0.35	0.69
Logistic Regression	0.69	0.67	**0.71**
Random Forests	**0.72**	0.88	0.70
eXtreme Gradient Boosting	0.70	**0.89**	0.70
Support Vector Machines	0.67	0.86	0.70

Table 3: Average results of multi-class classification

Summary

We present here a corpus of Bengali text for misinformation classification, structured along lines similar to prior work, comparable with English and Sinhala benchmarks cited herein. As with many languages in the Global South, data is relatively sparse, and we hope that the findings presented here will contribute to efforts in Bengali.

Classification performed using the same algorithms cited in said prior work reveals some interesting differences. While Sinhala work suggested that classifying CREDIBLE and UNCERTAIN might be a fruitful step, our results here show the greatest promise with CREDIBLE and FALSE - the binary that most such work uses. As with previous work, our results show promise for tree-based models (Random Forests and eXtreme Gradient Boosting here); as with the Sinhala work, Support Vector Machines (SVMs) are not as dominant as supposed.

There may be multiple reasons for this difference in performance. Plausible reason could be the differences between datasets, or the fundamental difference between linguistic structures affecting algorithms, as noted by Wijeratne & de Silva [4]. Further improvements may be made with deep learning, although we caution against deploying deep learning models with sparse data.

Acknowledgements

This research has been made possible through a grant from the Asia Foundation. We would like to thank Vihanga Jayawickrama and Asanka Ranasinghe for their assistance in this project, as well as the support of Ayesha Binti Towhid in the preliminary scoping work on misinformation in Bangladesh.

Appendix

List of sources for dataset

banglainsider.com
banglanews24.com
bd-pratidin.com
dailyinqilab.com
notuntvnews.com
bangla.asianetnews.com
bangladeshnewz.com
banglatribune.com
betanews24.com
coxsbazarlive24.com
daily-bangladesh.com
dainikpurbokone.net
dhakatimes24.com
ittefaq.com.bd
jagonews24.com
jugantor.com
kalerkantho.com
manobkantha.com.bd
mzamin.com
ntvbd.com
risingbd.com
samakal.com
sarabangla.net
somoynews.tv
youtube.com
ajkerjamalpur.com
banglanews24.com
chattogramdaily.com
coxsbazarnews.com
ctgcrimenews.com
ctgtimes.com
dainikamadershomoy.com
e-kantho24.com
lakshmipur24.com
poriborton.news
bdpress.agency
bn.mtnews24.com
prothomalo.com
1wwwbalerkontho.wordpress.com
ajkernewsgo.com
ajkernewz.com
ajkersatkhira.com
ajkersurjodoy24.com
amaderbrahmanbaria.com

amadercomillaa.com
amaderorthoneeti.com
amarcampus24.com
amarsangbad.com
archive.ph/1LJOg
atn24livenews.com
azviralnews.com
bagerhat24.com
bangla-bazaar.com
bangla.bdnews24.com
bangla.dhakatribune.com
bangla.thereport24.com
bangla24.com.bd
bangladesh24online.com
bangladesherkhabor.net
banglahunt.com
banglakatha.com
banglalivenews24.tv
banglarreporter.com
barta24.com
bbc-banglaa blogspot.my
bd-career.org
bdanalysis.com
bdnews24us.com
bdtime24.net
bekarjibon.coM
bhorer-dak.com
bhorerkagoj.com
bissoy.com
bograsangbad.com
boguracity.com
boishakhionline.com
businessbangladesh.com.bd
chandpur-kantho.com
chandpurtimes.com
channel24bd.tv
channelionline.com
ciencebee.com.bd
city24news.com
corporatesangbad.com
coxsbazarnews.com
cplusbd.net
dailyjanakantha.com
dailynayadiganta.com
dailynews23.com
dailysangram.com
dailysatkhira.com
dailysomoyersomikoron.com

dailysylhet.com
dailytnews.com
dainikamadershomoy.com
dainikpurbokone.com
daktarprotidin.com
deshreview.com
deshrupantor.com
dhakapost.com
dhakapostonline.com
djanata.com
easyreader.org
eisamay.indiatimes.com
ekushey-tv.com
enews71.com
etribune.net
eyenewsbd.com
Facebook.com
gazi24.com
Insafbd24.com
ishtiharnews.com
jagoreport24.com
jagoronnews.com
jamuna.tv
juystore.com
kaleralo.com
kalerkantho.com
lovenewsonline.com
mahcofficialnews
manobkantha.com.bdd
mohammadalijinnah.com
nayabarta.net
news.dailyekattor.com
news.priyo.com
news24bd.tv
newsbdplus.com
newsbhai24.com
newsjamuna.com
newskingbd.com
newsonline72.com
newzbigo.com
notuntvnews.comm
odhikar.news
onabil.net
onebd.news
onebd.news/
ourislam24.net
ppbd.news
ppbd.news/abroad

priyo.com
probasbd.com
prothomalo.com
provaterkhobor.com
publicvoice24.com
reatvender.com
report24today
rtvonline.com
rupali24bangla.com
Sahifa
sarabangla
satkhiranews.com
sebanews24.com
sheershakhobor.com
shikshabarta.com
shongbad.com.bd
silkcitynews.com
somewhereinblog.net
somoyerkonthosor.com
sonaymoritv.com
songinews24.com
sportspratidin.com
stvonlinenewsbd
sukhabor.com.bd
sunamganjbarta.com
swapnerkhulna.com
sylhetexpress.com
sylhettimes.net
sylhettoday24.news
thedaynightnews.com
unnews.com
viewer.com.bd
vinnonews.com
voiceofright.com
vornews.com
womeneye24.com
worldnewsaf.com
Youtube.com
zeenews.india.com

Bibliography

[1] Y. Wijeratne and D. C. Attanayake, "Artificial intelligence for text-based factchecking: Observations on the state and practicality of the art," 2021.

[2] V. Jayawickrama, A. Ranasinghe, D. C. Attanayake, and Y. Wijeratne, "A corpus and machine learning models for fake news classification in sinhala," 2021.

[3] L. Duong, "Natural language processing for resource-poor languages," Ph.D. dissertation, 2017.

[4] Y. Wijeratne, N. de Silva, and Y. Shanmugarajah, "Natural language processing for government: Problems and potential," *International Development Research Centre (Canada)*, 2019.

[5] Y. Wijeratne, "How much bullshit do we need? benchmarking classical machine learning for fake news classification," 2021.