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

Vihanga Jayawickrama, Asanka Ranasinghe, Dimuthu C. Attanayake, Yudhanjaya Wijeratne
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 3576 documents in Sinhala, drawn from Sri Lankan 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 Sinhala language, as well as comparisons to English benchmarks, and suggest that for smaller media ecosystems it may make more practical sense to model uncertainty instead of truth vs falsehood binaries.

Introduction

The detection of misinformation, or fake news, is a strong candidate for automation. As posited by Wijeratne et al (2021)[1], the velocity of misinformation is such that factcheckers may benefit from automated systems; as shown in repeated trials in the same paper, machine learning achieves a high degree of accuracy at using linguistic features to perform the kind of text classification required for this task. Nonetheless, a pressing issue in the field of natural language processing (NLP) is that a significant portion of research done is confined to resource rich languages such as English. Many languages, especially those that are confined to smaller geographical regions, endure a severe lack of resources that hinder the ability to truly exploit NLP and the tools it brings forth (Duong, 2017 [2], Wijeratne et al, 2019[3], Wijeratne et al, 2021[1]). This issue affects the Sinhala language as well. As de Silva (2019) [4] concludes, the little amount of NLP research that has been done on Sinhala stands as several isolated islands instead of a cohesive, continuous body of work. Little to no data exists around misinformation.

Thus we present an annotated dataset of Sinhala text for misinformation classification and the results of machine learning models to explore task fitness. Currently, no similar datasets are publicly available for the Sinhala language.

Methodology

Data acquisition and cleaning

Before processing, a primary dataset of 6417 news articles were scraped from 27 Sinhala news websites (Appendix A, Figure **??**), each website contributing not more than 200 documents. Only articles consisting of text were extracted, excluding articles consisting exclusively of multimedia in the process of scraping itself. Web scrapers were designed using Python for each of the websites to extract data.

Of the 6417 news articles, duplicates were first eliminated, and articles with a wordcount less that 20 (often found by manual examination to be either image captions) were removed. The remaining 6308 articles accounted for 98.02% of the original data scrape of news articles.

Each data point in this dataset consists 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

(Appendix B, Figure **??**) portrays the skew in dates in the dataset thus obtained. Most articles date from the end of 2020 to the beginning of 2021; however, some of the articles obtained from a few sites are published as far back in time as 2018, and date stamps of articles from mawbima.lk range within 3 days, from 11/3/2020 to 11/5/2020. This skew is reflective of both the frequency of publishing as well as errata in website structure that sometimes make it difficult to find conclusively dated articles.

From this, a randomized sample of 3000 articles was extracted for annotation. This we refer to as the primary dataset. Information regarding their sources, wordcounts and range of datestamps are provided in Appendix C.

In addition to the above, 576 pre-annotated documents were manually obtained from Sri Lanka-based factchecking operations Factcheck.lk, FactCrescendo.lk and Watchdog Sri Lanka. In collecting this data, only rumors judged to be false were collected; and of these, only the rumor content was collected, and not the explanation as to its veracity. This was concatenated into a secondary dataset for later integration.

Data annotation

The sample of 3000 articles from the primary dataset was divided equally among three individual annotators.A strict procedure was followed in the annotation process; each news article was categorized as "CREDIBLE," "FALSE," "PARTIAL," or "UNCERTAIN". The logic for the annotation schema a set of predetermined sources and a protocol, supplied by Watchdog Sri Lanka, were used in the verification process. This protocol, described below, relies on checking a tiered list of sources to see if the same information had been repeated or debunked.

Sources that had established themselves to have history of solid, factual reporting on Sri Lankan issues, a clearly visible demarcation between reporting of facts and personal opinions and little to no visible political bias were categorized as Tier 1. Legal documents, such as constitutions, government press releases and gazettes, also fall into Tier 1.

Sources that could be considered inconsistent (as regards news in Sri Lanka), reliable on certain topics but unreliable in others, were judged to be Tier 2. Sources considered to be downright polemical, or prone to exaggeration, publishing opinion disguised as fact, were judged Tier 3.

This tier list is the result of qualitative experience and not algorithmically determined.

In the verification process, each fact mentioned in a news article had to be verified adhering to the following 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

For an article to be annotated as CREDIBLE, FALSE or PARTIAL, a minimum of two Tier 1 sources or three Tier 2 sources were required to have published information that could be used to verify or disprove the contents of the article. Where only Tier 3 sources are available, a minimum of three sources was required for an article to earn a 'Credible' annotation, whereas a single higher-tier source was deemed sufficient to mark said article as False or Partial. If sufficient evidence could not be gathered to classify an article as CREDIBLE, FALSE, or PARTIAL, it would be classified as UNCERTAIN.

Once annotated, the accuracy of annotation was evaluated by a Watchdog Sri Lanka fact checker via three rounds of random sampling; where necessary, sampling was amended to reflect an overall acceptable error of no more than 5%.

To ensure better representation of data classes, the secondary dataset gathered from factcheckers was added to supplement the FALSE and PARTIAL categories. With the addition of them, the resultant annotated dataset included a total of 3576 documents: 1003 CREDIBLE, 568 FALSE, 118 PARTIAL, and 1887 UNCERTAIN.

Category	**Article Count**		
	From News Websites	**From factcheckers**	**Total**
CREDIBLE	1003	0	1003
FALSE	27	541	568
PARTIAL	83	35	118
UNCERTAIN	1887	0	1887
Total	**3000**	**576**	**3576**

Table 1: Breakdown of articles in the annotated dataset

Machine learning for task fitness

Wijeratne et al (2021)[5] demonstrated that classical machine learning can achieve relatively high degrees of accuracy with a binary categorization of classes; usually RELIABLE and FAKE or TRUE and FALSE. The data range we have collected here lies within the lower bounds of data range tested there.

To test such a classification with the closest equivalent in our data, a reduced dataset was created, of 568 articles each from the CREDIBLE and FALSE categories. This dataset is class-balanced. All data was cleaned of punctuation and all other non-Sinhala characters, as well as stop words. The set of stop words has been derived from the corpus itself following the procedure mentioned in Wijeratne & de Silva (2020) [6]. The algorithms mentioned in our previous benchmarks [5] - 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	0.792	0.631
Logistic Regression	0.822	0.801
Random Forests	**0.841**	**0.825**
eXtreme Gradient Boosting	0.828	0.819
Support Vector Machines	0.832	0.817

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

In a Sri Lankan context, such a ready classification between CREDIBLE and FALSE may not be easy to establish. Relative to the anglosphere on which [1] tested algorithms on, a much smaller media ecosystem, low media freedom and tight political links among media ownership[1] mean that a relative lack of reliable sources. The dataset we have collected confirms this assesment, as the representation of overtly FALSE news from the primary dataset is extremely low (27). Indeed, the job of factcheckers involves conducting investigative journalism on suspicious news to establish whether they are credible or false. It may therefore be more useful to model CREDIBLE vs UNCERTAIN, using the same 568 document-count per class:

Model Used	**F1 Score**	
	Credible	**Uncertain**
Naive Bayes	0.598	0.764
Logistic Regression	0.817	0.817
Random Forests	**0.854**	0.832
eXtreme Gradient Boosting	**0.854**	**0.839**
Support Vector Machines	0.843	0.829

Table 3: 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	0.521	0.162	0.550
Logistic Regression	0.705	0.798	0.798
Random Forests	**0.725**	0.801	**0.833**
eXtreme Gradient Boosting	0.705	0.791	0.808
Support Vector Machines	0.711	**0.808**	0.817

Table 4: Average results of multi-class classification

[1] https://sri-lanka.mom-rsf.org/en/

Summary

Firstly, we present a corpus of Sinhala misinformation, suited for further work in misinformation and media studies, automated classification efforts and for discourse analysis. Natural Language Processing in Sinhala is still at an elementary stage, owing to the lack of resources associated with the language ([4]. We hope the findings presented through this paper would contribute to that foundation, upon which researchers can continue to build intriguing castles.

Secondly, classification work using the same algorithms presented in Wijeratne et al (2021)[5] reveals some interesting differences. It may hold true that for many countries in similar situations, where the media ecosystem is relatively small, a lack of sources may make it difficult to build good machine learning datasets that have clear demarcations between true and false. The previous paper asked what practical deployment of automated factchecking models might require, pointing out that compute and data costs need be factored in. We would add that it may also be wise to consider shifting the goalposts and modelling uncertainty instead of rigid binaries.

Modelling an inherently uncertain class also introduces interesting epistemological challenges that may bleed over into the issue of noise in data. Our results for CREDIBLE vs UNCERTAIN show promise for tree-based models (Random Forests and eXtreme Gradient Boosting here), with F1 scores in both classes over the 83% mark for accuracy, although whether the general machine learning paradigm of 'more data = better' will hold true is yet to be seen.

Thirdly, contrary to much text classification work where Support Vector Machines (SVMs) consistently score the highest accuracy, here Random Forests outperform other options. [5] proposed that SVMs be used in instances of low training data, owing to the nature of their fit time. However, here SVMs appear at a minute disadvantage.

There may be multiple reasons for this. Plausible reason could be the differences between datasets, or the fundamental difference between linguistic structures affecting algorithms, as noted by Wijeratne & de Silva [3]. Nevertheless, results are promising. Additional gains may be had by employing deep learning; preliminary testing yielded F1 scores of >0.7 and >0.8 with a 3 and 5-layer neural network respectively, although we have chosen not to include these results above.

Acknowledgements

This research has been made possible through a grant from the Asia Foundation. One of the authors of this paper, Yudhanjaya Wijeratne, is a co-founder of Watchdog Sri Lanka, which supplied contextual information and aided in the annotation process and training annotators.

Appendix A

Website	URL	Article Count
ada.lk	https://www.ada.lk	200
adaderana.lk	http://sinhala.adaderana.lk	200
anidda.lk	https://www.anidda.lk	200
aruna.lk	http://www.aruna.lk	429
asianmirror.lk	https://am.lk	200
bbc.com	https://www.bbc.com/sinhala	286
deshaya.lk	https://www.deshaya.lk	299
dinamina.lk	http://www.dinamina.lk	200
divaina.com	https://divaina.com	314
gossiplankanews.com	https://www.gossiplankanews.com	200
gossip-lankanews.com	https://www.gossip-lankanews.com	450
gossip.hirufm.lk	https://gossip.hirufm.lk	429
hirunews.lk	https://www.hirunews.lk	270
lankacnews.com	https://lankacnews.com	200
lankadeepa.lk	https://www.lankadeepa.lk	200
lankaenews.lk	http://www.lankaenews.lk	200
mawbima.lk	https://mawbima.lk	200
nethnews.lk	http://nethnews.lk	200
newsfirst.lk	https://www.newsfirst.lk	200
newshub.lk	https://newshub.lk	200
praja.lk	http://praja.lk	124
ravaya.lk	https://ravaya.lk	309
roar.media	https://roar.media	200
samabima.com	https://www.samabima.com	269
sarasaviya.lk	http://www.sarasaviya.lk	200
silumina.lk	http://www.silumina.lk	38
vikalpa.org	https://www.vikalpa.org	200

Number of articles scraped from each source

Appendix B

Website	Date Range
ada.lk	29/10/2020 - 04/11/2020
adaderana.lk	28/10/2020 - 04/11/2020
anidda.lk	06/05/2020 - 03/11/2020
aruna.lk	25/10/2020 - 04/11/2020
asianmirror.lk	28/10/2020 - 05/11/2020
bbc.com	22/11/2019 - 04/11/2020
deshaya.lk	24/04/2020 - 31/10/2020
dinamina.lk	30/10/2020 - 05/11/2020
divaina.com	07/09/2020- 04/11/2020
gossiplankanews.com	29/10/2020 - 04/11/2020
gossip-lankanews.com	25/09/2020 - 04/11/2020
gossip.hirufm.lk	17/10/2020 - 04/11/2020
hirunews.lk	29/10/2020 - 04/11/2020
lankacnews.com	26/10/2020 - 05/11/2020
lankadeepa.lk	30/08/2020 - 04/11/2020
lankaenews.lk	07/02/2020- 27/10/2020
mawbima.lk	03/11/2020 - 05/11/2020
nethnews.lk	25/10/2020 - 04/11/2020
newsfirst.lk	25/10/2020 - 11/04/2020
newshub.lk	30/10/2020 - 05/11/2020
praja.lk	12/12/2018 - 21/10/2020
ravaya.lk	29/05/2020 - 29/10/2020
roar.media	12/11/2018 - 21/10/2020
samabima.com	12/07/2019 - 02/11/2020
sarasaviya.lk	06/08/2020 - 05/11/2020
silumina.lk	02/11/2020 - 07/11/2020
vikalpa.org	04/04/2020 - 04/11/2020

Date range covered by each source in the original dataset

Appendix C

Website	Date Range
ada.lk	29/10/2020 - 04/11/2020
adaderana.lk	28/10/2020 - 04/11/2020
anidda.lk	06/05/2020 - 03/11/2020
aruna.lk	25/10/2020 - 04/11/2020
asianmirror.lk	28/10/2020 - 05/11/2020
bbc.com	27/11/2019 - 03/11/2020
deshaya.lk	24/04/2020 - 31/10/2020
dinamina.lk	30/10/2020 - 05/11/2020
divaina.com	07/09/2020- 04/11/2020
gossiplankanews.com	29/10/2020 - 04/11/2020
gossip-lankanews.com	25/09/2020 - 04/11/2020
gossip.hirufm.lk	17/10/2020 - 04/11/2020
hirunews.lk	29/10/2020 - 04/11/2020
lankacnews.com	26/10/2020 - 05/11/2020
lankadeepa.lk	30/08/2020 - 03/11/2020
lankaenews.lk	11/02/2020- 26/10/2020
mawbima.lk	03/11/2020 - 05/11/2020
nethnews.lk	25/10/2020 - 04/11/2020
newsfirst.lk	25/10/2020 - 11/04/2020
newshub.lk	30/10/2020 - 05/11/2020
praja.lk	12/12/2018 - 05/09/2020
ravaya.lk	29/05/2020 - 29/10/2020
roar.media	12/11/2018 - 21/10/2020
samabima.com	18/07/2019 - 02/11/2020
sarasaviya.lk	06/08/2020 - 05/11/2020
silumina.lk	02/11/2020 - 07/11/2020
vikalpa.org	04/04/2020 - 04/11/2020

Date range covered by each source in the original dataset

Bibliography

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

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

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

[4] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research," *arXiv preprint arXiv:1906.02358*, 2019.

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

[6] Y. Wijeratne and N. de Silva, "Sinhala language corpora and stopwords from a decade of sri lankan facebook," *arXiv preprint arXiv:2007.07884*, 2020.