NSina: A News Corpus for Sinhala

Hansi Hettiarachchi[1], Damith Premasiri[2], Lasitha Uyangodage[3], Tharindu Ranasinghe[4]
[1]Birmingham City University, UK, [2]Lancaster University, UK,
[3]University of Münster, Germany, [4]Aston University, UK
[1]hansi.hettiarachchi@bcu.ac.uk, [2]d.dolamullage@lancaster.ac.uk,
[3]luyangod@uni-muenster.de, [4]t.ranasinghe@aston.ac.uk

Abstract

The introduction of large language models (LLMs) has advanced natural language processing (NLP), but their effectiveness is largely dependent on pre-training resources. This is especially evident in low-resource languages, such as Sinhala, which face two primary challenges: the lack of substantial training data and limited benchmarking datasets. In response, this study introduces NSina, a comprehensive news corpus of over 500,000 articles from popular Sinhala news websites, along with three NLP tasks: news media identification, news category prediction, and news headline generation. The release of NSina aims to provide a solution to challenges in adapting LLMs to Sinhala, offering valuable resources and benchmarks for improving NLP in the Sinhala language. NSina is the largest news corpus for Sinhala, available up to date.

Keywords: news corpus, low-resource languages, text classification, text generation

1. Introduction

The recent emergence of large language models (LLMs) has ushered in significant advancements in the field of natural language processing (NLP) (Devlin et al., 2019). These LLMs have produced state-of-the-art results in many NLP benchmarks, outperforming previous machine learning models such as LSTMs (Lin et al., 2022). Beyond academic research, these recent LLMs, such as GPT (Brown et al., 2020), have also driven the development of widely popular products, including chatbots, machine translation and writing assistants, among other applications, making them highly popular among the general public.

While LLMs have achieved considerable success and garnered popularity, their effectiveness heavily relies on access to resources for weight pre-training. Consequently, these LLMs excel in high-resource languages but encounter challenges when applied to low-resource languages (Wang et al., 2020). Two primary factors contribute to the complexities of deploying LLMs in low-resource linguistic contexts: (1) Limited availability of open-access corpora for pre-training the models in low-resource languages. and (2) The absence of appropriate benchmarking datasets in these languages to assess the performance of the models. In this research, we address these challenges for Sinhala by releasing NSina: A large News Corpus for Sinhala accompanied by a range of benchmarking tasks.

Sinhala is an Indo-Aryan language spoken by over 17 million people in Sri Lanka. Sinhala is one of the two official languages in Sri Lanka. Predominantly, the Sinhala-speaking community comprises the Sinhalese people, the largest ethnic group on the island. Despite its sizable community, Sinhala remains relatively under-resourced compared to other languages spoken in the region (De Silva, 2019). The complexities regarding LLMs and low-resource languages that we discussed previously appear in Sinhala too. While there exist several multilingual language models, such as XLM-RoBERTa (Conneau et al., 2020), which offer support for Sinhala, the multilingual corpora used to train these models contain a relatively limited proportion of Sinhala compared to other languages (Wang et al., 2020). For example, the OSCAR 23.01 multilingual corpus (Abadji et al., 2022), which has been used to train multilingual language models, only contains 2.6GB Sinhala text which does not even contribute to 1% of the total dataset. Furthermore, since these data were automatically extracted from Common Crawl dumps, they contain noise resulting from boilerplate content extracted from headers, footers and sidebars of web crawls (Abadji et al., 2022).

There is also a pre-trained BERT model trained specifically for Sinhala (Dhananjaya et al., 2022). However, it's worth noting that this model has been trained on a relatively small Sinhala corpus, leading to its inability to consistently outperform multilingual LLMs in various Sinhala NLP tasks, as evidenced in both Dhananjaya et al. (2022) and Ranasinghe et al. (2024). These constraints primarily stem from the limited availability of sizable Sinhala corpora for model training. Furthermore, as we mentioned before, there is a huge limitation of available benchmarking datasets for Sinhala. This is also evident in Dhananjaya et al. (2022), where they evaluated the pre-trained BERT model in only three text classification tasks.

In this paper, we lay the foundation to address these challenges. Firstly, we compiled a large news corpus with more than 500,000 news articles from ten popular news websites in Sri Lanka. Secondly,

LREC-COLING 2024, pages 12307–12312

we compose three NLP tasks from the news corpus, including two text classification tasks: (1) News media identification (2) News category prediction and one text generation task: (3) News headline generation. We release separate train and test sets for each task, which are sampled from NSINA. These datasets can be used as benchmarks to evaluate the LLMs in Sinhala. While there are several news corpora for Sinhala (Upeksha et al., 2015), NSINA is the most updated and the largest Sinhala news corpus available.

Our **main contributions** are;

1. We introduce NSINA, a large News Corpus for Sinhala, which comprises 506,932 news articles, and we describe the steps taken to compile it.
2. We introduce three NLP tasks associated with NSINA. (i) News media identification, (ii) News category prediction, and (iii) News headline generation. We release train and test sets for each task, which are subsampled from NSINA.
3. We evaluate several machine learning models on each task and compare the performance.
4. We release NSINA, as an open-access publicly available dataset alongside the trained machine-learning models for each task[1].

2. Dataset Construction

We first present the data collection methodology we used followed by a detailed statistical analysis of NSINA.

2.1. Data Collection

First, we identified ten news sources that are popular in Sri Lanka. These sources encompass a diverse range, including "Adaderana," "ITN News," "Lankatruth," "Divaina," "Hiru News," "Lankadeepa," "Vikalpa," "Dinamina," and "Siyatha." Importantly, we took deliberate steps to ensure a balanced representation within this selection, encompassing both pro-government and anti-government news outlets.

We used a Python script to scrape each news media website. For each news article, we extracted the source, timestamp, headline, news content, URL and category.

2.2. NSINA

After data collection, we performed data cleaning. There were a considerable amount of news articles that had less than ten Sinhala words. These articles were identified and filtered. The final dataset consisted of 506,932 news articles. The source news media of these news articles is shown in Table 1. As can be seen, "Lankadeepa" and "Hiru News" contribute most to the corpus, having more than 100,000 articles each.

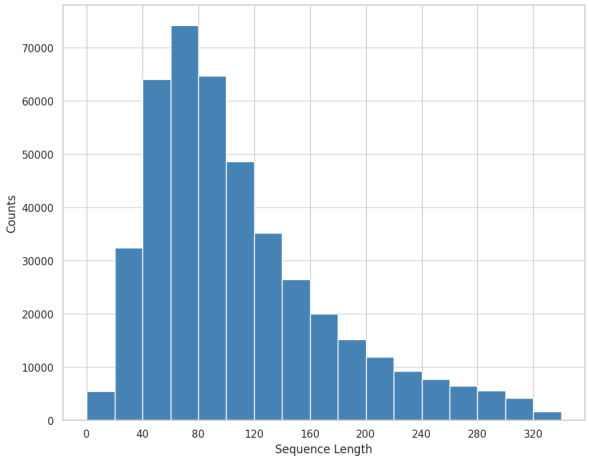

Figure 1: Token frequency distribution of news content in NSINA

Source	Amount
Adaderana	83918
ITN News	30777
Lankatruth	48180
Divaina	26043
Hiru News	130729
Sinhala News LK	20371
Lankadeepa	141663
Vikalpa	14309
Dinamina	7642
Siyatha	3300
Total	506932

Table 1: Number of news articles from each source in NSINA.

Figure 1 shows the token frequency distribution of the news content after removing the outliers (very long documents). It is clear that most news articles contain around 60-120 tokens. Furthermore, in Figure 2, we plot the token frequency of headlines. The figure shows that most of the titles are short, having between 6-12 tokens. The corpus has more than 1 million tokens and more than 100,000 unique tokens.

Previous Sinhala news corpus; SinMin (Upeksha et al., 2015) was only 1.01 GB in size. Compared to that, NSINA is 1.87 GB and provides a larger and updated news corpus. Furthermore, the Sinhala portion of the OSCAR 23.01 corpus (Abadji et al., 2022) only has 301,066 documents and compared to that NSINA has more than 500,000 properly cleaned Sinhala documents.

[1]The dataset is available at https://github.com/Sinhala-NLP/NSINA

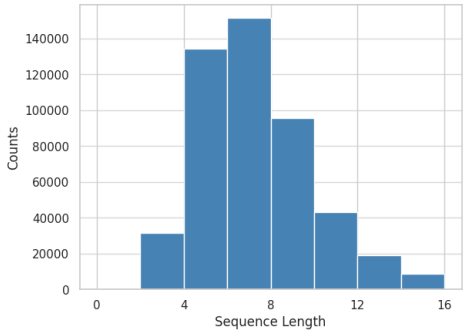

Figure 2: Token frequency distribution of news headline in NSina

3. Tasks

We compiled the following three tasks from NSina. We believe that creating benchmarks would help to evaluate the LLMs in Sinhala NLP tasks.

1. News media identification
2. News category prediction
3. News headline generation

Each of the tasks had different train/ test sets and machine learning models. The following subsections will describe tasks, machine learning models trained for the task and the results.

3.1. News Media Identification

The first task we compiled from NSina is identifying news source given the news content. This task serves a dual purpose: firstly, it aids in recognising the distinctive style of each news outlet in their news presentation. Moreover, this task can be further explored to unveil potential political biases within Sri Lankan news media. This is a text classification task, where the input to the ML models would be the news content, and the expected output of the model is the news media name.

3.1.1. Data

As all the news instances in NSina contained its news source, constructing the train/ test set for this task was straightforward. However, as can be seen in Table 1, since some news media sources have more instances, we undersampled them. We only used 10,000 instances from each news source. For the two sources that had less than 10,000 instances ("Dinamina" and "Siyatha") we used the original number of instances they contained. We divided this dataset into a training and test set following a 0.8 split[2].

[2]The sampled dataset is available at https://github.com/Sinhala-NLP/Sinhala-News-Media-Identification

3.1.2. Models

For this task, we experimented with several text classification models based on transformers. From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = softmax(W\boldsymbol{h}+b)$, where $W \in \mathbb{R}^{k\times d}$ is the softmax weight matrix, and k is the number of labels.

For these, we employed a batch-size of 16, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were evaluated at least three times per epoch while training using an evaluation set that had one-fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs.

3.1.3. Results

We evaluate the models using the average weighted F1 score and macro F1 score. The results are shown in Table 2.

Models	Weighted F1	Macro F1
`XLM-R Large`	0.8917	0.8763
`XLM-R Base`	0.8961	0.8854
`SinBERT Small`	0.8966	0.8883
`SinBERT Large`	**0.8967**	**0.8895**

Table 2: News media classification results. The results are ordered with the ascending order for **Macro F1** score.

As can be seen in the results, the performance of all the transformer models on the task was excellent. All the models achieved more than 0.88 Macro F1 score on the task. These outcomes imply that each news source exhibits a distinct stylistic approach in presenting their news content.

Notably, the `SinBERT Large` model outperformed the `XLM-R Large` model. However, the `XLM-R Large` model produces very close results to the `SinBERT Large` model.

3.2. News Category Prediction

The second task we compiled from NSina was news category prediction. This is an old task in NLP, but it has many potential applications (Bracewell et al., 2009). Given the news content, the ML models should predict a pre-defined category for the news.

3.2.1. Data

Preparing the train/test sets for this task was more challenging than the previous task as the categories are not commonly defined for all news media. First, for this task, we dropped all the news articles without a category as some news sources prefer not to categorise them. Next, we identified the top 100 news categories from the available news instances. We grouped them into four main categories: local news, international news, sports news, and business news. To avoid bias, we undersampled the dataset. We only used 10,000 instances from each category, and for the "Business" category, we used the original number of instances which was 8,777 articles. We divided this dataset into a training and test set following a 0.8 split[3].

3.2.2. Models

We employed the same ML models experimented with for the previous task, as this is also a text classification task. The hyperparameter configurations and training strategy also remained the same.

3.2.3. Results

Similar to the previous task, we used Weighted F1 and Macro F1 scores to evaluate the models. The results are shown in Table 3.

Models	Weighted F1	Macro F1
SinBERT Small	0.9310	0.9306
SinBERT Large	0.9370	0.9367
XLM-R Large	0.9385	0.9381
XLM-R Base	**0.9387**	**0.9382**

Table 3: News category prediction results. The results are ordered with the ascending order for **Macro F1** score. The best results are in bold.

Similar to the previous task, transformer models provided excellent results in this task as well. This can be a particularly easy task as the news content within each category we analysed is distinct and varied. Notably, XLM-R models could outperform SinBERT model in this task.

3.3. News Headline Generation

Lastly, we created a natural language generation (NLG) task using the NSina corpus, where the ML model's objective is to generate news headlines based on the provided news content. As all the news content in NSina has headlines, constructing this NLG task was straightforward. In an era of increasing interest in language generation models such as GPT, this benchmark will be invaluable for evaluating their performance in the context of Sinhala language generation.

3.3.1. Data

We used the same instances from NSina as all the news articles had headlines. We divided this dataset into a training and test set following a 0.8 split[4].

3.3.2. Models

In this task, we exploit two types of ML models based on transformers.

General Transformers - We created a Seq2Seq model from general transformers by adding a transformer decoder, which takes the encoder's output and generates the target sequences. We experimented with several general-purpose transformer models that support Sinhala, including XLM-Roberta (Conneau et al., 2020), and SinBERT (Dhananjaya et al., 2022).

Text Generation Transformers - We also experimented with several text generation transformers as they have provided excellent results in text generation tasks. Specifically, we explored mBART (Lewis et al., 2020) and several mT5 (Xue et al., 2021) variants. Both mBART and mT5 support Sinhala.

For both types of transformer models, we employed a batch size of 16, Adam optimiser with learning rate $1e{-}4$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model were updated. The models were trained only using the training data and evaluated while training using an evaluation set that had one-fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs.

3.3.3. Results

We evaluated the results of the models using two popular NLG evaluation metrics: BLEU (Papineni et al., 2002) and Translation Edit Rate (TER). While there are advanced NLG metrics such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2019), they do not currently support Sinhala.

The results in Table 4 show that all the transformer models do not perform well in this natural

[3]The sampled dataset is available at https://github.com/Sinhala-NLP/Sinhala-News-Category-Prediction/

[4]The sampled dataset is available at https://github.com/Sinhala-NLP/Sinhala-Headline-Generation

Models	BLEU	TER
SinBERT	0.08	0.80
XLM-R Base	0.11	0.76
XLM-R Large	0.14	0.75
mBART	0.15	0.74
mT5 Base	0.16	0.73
mT5 Large	0.17	0.72

Table 4: News category prediction results. The results are ordered in ascending order for BLEU.

language processing task. The mT5-large model produced the best result with a BLEU score of 0.17. However, a BLEU score in the range of 0.1 and 0.2 suggests that there is a poor overlap between the generated headline and the actual headline.

The poor performance on NLG can be attributed to two main reasons. (1) There is no model trained specifically for Sinhala language generation and (2) The BLEU and TER scores cannot properly evaluate Sinhala text generation and there is a need for advanced NLG metrics for Sinhala.

4. Conclusion

This research presented NSina, a large news corpus for Sinhala that can be used to train LLMs. NSina is larger and more recent than previous news corpus released for Sinhala, such as Sin-Min (Upeksha et al., 2015). We also release three benchmark datasets sampled from NSina that can be used to evaluate LLMs. We evaluated several transformer models on each task. The results show that multilingual transformer models such as XLM-R provide very close results or sometimes even outperform language-specific models such as Sin-BERT (Dhananjaya et al., 2022), suggesting more research should be done to train Sinhala-specific transformer models. Furthermore, all the experimented models perform poorly on the proposed NLG task, suggesting that more language generation models should be explored for Sinhala.

In future work, we would like to utilise NSina with other available Sinhala resources to create robust transformer models. Furthermore, a GLUE (Wang et al., 2018) like benchmark will be created for Sinhala, including the tasks proposed in this paper, which will serve as a pivotal platform for evaluating the proficiency of LLMs in processing Sinhala.

Acknowledgments

The computational experiments in this paper were conducted on the Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

Ethics Statement

NSina was collected from publicly available websites, and none of the records were edited in the process. Similar to the previous research that compiled and released news corpora (Nagoudi et al., 2020; Kakwani et al., 2020), for every instance in NSina, we released the URL for the original news article. Furthermore, we released NSina with the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 International Public License, which prevents users from altering the instances in the dataset. While NSina is publicly available in HuggingFace, we only released it as a gated dataset where the users need to accept the license and request access. All the datasets released for the subtasks in this paper also follow a similar process.

5. Bibliographical References

David B Bracewell, Jiajun Yan, Fuji Ren, and Shingo Kuroiwa. 2009. Category classification and topic discovery of japanese and english news articles. *Electronic Notes in Theoretical Computer Science*, 225:51–65.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

6. Language Resource References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Tariq Alhindi. 2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. SOLD: Sinhala offensive language dataset. *Language Resources and Evaluation*.

Dimuthu Upeksha, Chamila Wijayarathna, Maduranga Siriwardena, Lahiru Lasandun, Chinthana Wimalasuriya, NHND De Silva, and Gihan Dias. 2015. Implementing a corpus for sinhala language. In *Symposium on Language Technology for South Asia 2015*.