Sinhala Encoder-only Language Models and Evaluation

Tharindu Ranasinghe[1], **Hansi Hettiarachchi**[1], **Nadeesha Pathirana**[2], **Damith Premasiri**[1], **Lasitha Uyangodage**[3], **Isuri Anuradha**[1], **Alistair Plum**[4], **Paul Rayson**[1] **and Ruslan Mitkov**[1]
[1]School of Computing and Communications, Lancaster University, UK
[2] School of Computer Science and Digital Technologies, Aston University, UK
[3] University of Münster, Germany [4] University of Luxembourg, Luxembourg
t.ranasinghe@lancaster.ac.uk

Abstract

Recently, language models (LMs) have produced excellent results in many natural language processing (NLP) tasks. However, their effectiveness is highly dependent on available pre-training resources, which is particularly challenging for low-resource languages such as Sinhala. Furthermore, the scarcity of benchmarks to evaluate LMs is also a major concern for low-resource languages. In this paper, we address these two challenges for Sinhala by (*i*) collecting the largest monolingual corpus for Sinhala, (*ii*) training multiple LMs on this corpus and (*iii*) compiling the first Sinhala NLP benchmark (SINHALA-GLUE) and evaluating LMs on it. We show that the Sinhala LMs trained in this paper outperform the popular multilingual LMs, such as XLM-R and existing Sinhala LMs in downstream NLP tasks. All the trained LMs are publicly available. We also make SINHALA-GLUE publicly available as a public leaderboard, and we hope that it will enable further advancements in developing and evaluating LMs for Sinhala.

1 Introduction

The recent developments of language models (LMs) have shown significant advancements in the field of natural language processing (NLP) (Devlin et al., 2019) as they have produced state-of-the-art results in many NLP tasks, outperforming previous machine learning models such as LSTMs (Lin et al., 2022). Various language understanding benchmarks like GLUE (Wang et al., 2018) and SUPERGLUE (Wang et al., 2019) have been created to evaluate and compare these LMs. Successful LMs have been deployed widely in NLP applications such as machine translation (Haddow et al., 2022; Xu et al., 2024), chatbots (Adamopoulou and Moussiades, 2020; Zheng et al., 2023), and writing assistants (Min et al., 2023; Kobayashi et al., 2024), which have gained significant popularity among the general public (Yao et al., 2024).

Although LMs have attained notable success and widespread popularity, their effectiveness largely depends on access to language resources for model pre-training (Shikali and Mokhosi, 2020). Multilingual language models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) have tried to address the resource-scarcity of low-resource languages through techniques like cross-lingual transfer learning (Artetxe et al., 2020). However, due to the small data sizes of low-resource languages, subword tokenisers trained jointly on multiple languages tend to over-split the tokens of such languages, and LMs are not able to learn good quality representations of them (Hangya et al., 2022; Wu and Dredze, 2020; Rust et al., 2021). As a result, models trained exclusively on a single language have demonstrated superior performance on downstream tasks in the corresponding language compared to their multilingual counterparts (Straka et al., 2021). In response to this limitation, the NLP community has released numerous monolingual LMs tailored to individual languages (Koutsikakis et al., 2020; Nguyen and Tuan Nguyen, 2020; Cañete et al., 2022).

Sinhala, an Indo-Aryan language, is spoken by more than 17 million people in Sri Lanka and is recognised as one of the nation's two official languages. Predominantly, the Sinhalese community, the largest ethnic group in Sri Lanka, constitutes the bulk of Sinhala speakers. Despite its significant number of users, Sinhala is relatively under-resourced compared to other languages in the region (De Silva, 2019). According to Joshi et al. (2020), Sinhala is classified in the group of '*Left-Behinds*'; a group of languages that has been largely neglected in the development of language technologies. The authors conclude that lifting such languages up in the digital space will be a monumental, probably impossible effort due to the severe scarcity of linguistic resources (Joshi et al., 2020).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8623–8636
July 27 - August 1, 2025

While several multilingual language models, such as XLM-RoBERTa (Conneau et al., 2020) and info-XLM (Chi et al., 2021), support Sinhala, the multilingual datasets used to train these models allocate only a modest share to Sinhala compared to other languages (Wang et al., 2020). For instance, OSCAR 23.01 (Abadji et al., 2022), which is used to train these multilingual models, comprises just 2.6GB of Sinhala text, contributing to less than 1% of the total dataset. A dedicated Sinhala BERT model has also been developed (Dhananjaya et al., 2022), but its training is constrained by the relatively small size of the available Sinhala corpus. As a result, its performance does not consistently surpass that of multilingual LMs across various Sinhala NLP tasks, as demonstrated in previous studies (Dhananjaya et al., 2022; Ranasinghe et al., 2024a; Hettiarachchi et al., 2024). These limitations stem primarily from the scarcity of large-scale Sinhala corpora for training.

Furthermore, as we mentioned before, there is a significant research gap in the available benchmarking datasets for Sinhala. Ranathunga and de Silva (2022) report that only 1.14% of Sinhala NLP papers have released the relevant data sets in public repositories. Therefore, a GLUE like benchmark is crucial for Sinhala. This is also evident in Dhananjaya et al. (2022), where the pre-trained Sinhala BERT model is evaluated only on three text classification tasks.

In this paper, we address these research gaps in Sinhala NLP with the following **main contributions**.

(*i*) We **collect and release the largest monolingual corpus** for Sinhala, which can be used to train Sinhala LMs.

(*ii*) We **train three different monolingual pre-trained transformer models** on this corpus that support Sinhala, amounting to the largest collection of transformers available in Sinhala.

(*iii*) We **compile the first language understanding benchmark in Sinhala**; SINHALA-GLUE with nine NLP tasks. We evaluate the pre-trained transformer models that we trained in *(ii)* with the already available Sinhala transformer models (SinBERT (Dhananjaya et al., 2022) and multilingual LMs that support Sinhala, such as XLM-RoBERTa (Conneau et al., 2020) and Info-XLM (Chi et al., 2021). We show that the models introduced in this paper outperform the multilingual and previous Sinhala LMs.

2 Related Work

2.1 Sinhala Natural Language Processing

Sinhala is the native language of the Sinhalese people, the largest ethnic group in Sri Lanka. It belongs to the vast Indo-European language family. As we mentioned before, despite the large speaker base, Sinhala remains a low-resource language in the NLP world. The scarcity of annotated datasets makes it particularly challenging to evaluate language models effectively.

Addressing these gaps, multiple NLP datasets have been released for Sinhala in the last few years, including offensive language detection (Ranasinghe et al., 2024a), sentiment analysis (Ranathunga and Liyanage, 2021), headline generation (Hettiarachchi et al., 2024), machine translation (Pushpananda et al., 2024) and text summarisation (Hewapathirana et al., 2024). For a more detailed survey, we refer the authors to De Silva (2019), which has been updated frequently.

Several multilingual LMs, such as XLM-R, support Sinhala. However, due to Sinhala's own unique writing system derived from the Indian Brahmi script (Bandara et al., 2012), the majority of the subword tokenisers trained jointly in multiple languages over-split Sinhala words (Velayuthan and Sarveswaran, 2025). Therefore, multilingual models provide sub-optimal results in some Sinhala NLP tasks (Shardlow et al., 2024). The lack of an evaluation benchmark has made it challenging to have broader conclusions.

As we mentioned before, a Sinhala BERT model also exists (Dhananjaya et al., 2022), which has been trained on a rather small corpus followed by a limited evaluation. A few studies have highlighted that multilingual models outperform the Sinhala BERT model in several NLP tasks (Hettiarachchi et al., 2024; Ranasinghe et al., 2024a).

2.2 NLU Benchmarks

The GLUE benchmark comprises 11 natural language understanding (NLU) tasks, including semantic textual similarity, natural language inference, and various classification challenges (Wang et al., 2018). Subsequently, this benchmark was expanded to include more advanced and complex tasks in its SUPERGLUE version (Shavrina et al., 2020). Both GLUE and SUPERGLUE are restricted to English.

Several benchmarks have been introduced to support the development and evaluation of models in other languages. When they are categorised by the language family, for the *Sino-Tibetan* family, both CLUE (Xu et al., 2020) and CUGE (Yao et al., 2021) focus on Chinese. In the *Romance* family, benchmarks have been developed for French (Le et al., 2020), Italian (Basile et al., 2023) and Catalan (Armengol-Estapé et al., 2021). The *Balto-Slavic* group, benchmarks includes Russian (Shavrina et al., 2020), Bulgarian (Hardalov et al., 2023) and Slovenian (Žagar and Robnik-Šikonja, 2022). The *Altic* language group includes Korean (Park et al., 2021), while the *Iranian* family includes Persian (Khashabi et al., 2021).

Recently, several multilingual benchmarks have also been developed. Liang et al. (2020) proposed XGLUE, a benchmark for 19 languages that covers NLP tasks such as named entity recognition, news classification and headline generation. Hu et al. (2020) collected a cross-lingual evaluation dataset in 40 languages, later extended with 10 additional (Ruder et al., 2021), including tasks similar to the SUPERGLUE setup including token classification, question answering and textual similarity. However, none of these benchmarks include Sinhala and therefore, it has been challenging to evaluate language models in Sinhala. In this paper, we address this challenge by introducing SINHALA-GLUE.

3 SINHALA-Corpus1.5B: Sinhala Monolingual Corpus

We gathered Sinhala textual data from diverse sources, including web articles, news media, social media, books and government documents, utilising six openly available datasets to create the Sinhala monolingual corpus. As summarised in Table 1, it contains over 1.5 billion tokens across more than 3.5 million documents.

HPLT 2.0 (de Gibert et al., 2024) is a multilingual corpus extracted from the Internet Archive and Common Crawl, covering 75 languages, including Sinhala. (License: CC0)

FineWeb2 (Penedo et al., 2024) is the upgraded version of the FineWeb dataset, including text data for over 1,000 languages, collected from 96 CommonCrawl snapshots from 2013 to 2024. It includes a Sinhala subset, ranking among the top 80 languages by data size. (License: ODC-By 1.0)

NSina (Hettiarachchi et al., 2024) is a comprehensive collection of news articles from ten Sinhala news websites popular in Sri Lanka. These sources encompass both pro- and anti-government news outlets, ensuring a balanced representation. (License: CC BY-NC-SA 4.0)

FacebookDecadeCorpora (FDC) (Wijeratne and de Silva, 2020) is a social media corpus extracted from Sri Lankan Facebook pages, spanning 2010 to 2020. It covers data from diverse categories, including politics, media and celebrities. (License: CC BY 4.0)

SinMin (Upeksha et al., 2015) is an extensive Sinhala corpus composed of modern and old texts of different genres and styles. Its primary sources include online newspapers and magazines, school textbooks, Mahawansa (the historical chronicle of Sri Lanka), Sinhala Wikipedia, Sri Lankan gazette and Sinhala subtitles. (License: CC BY)

SemiSOLD (Ranasinghe et al., 2024a) is a large collection of Sinhala tweets, initially extracted to create an offensive language detection dataset for Sinhala. The tweets were labelled for offensive content, and only the non-offensive ones were included in the Sinhala corpus. (License: CC BY 4.0)

Dataset	#Tokens	#Documents	Disk Size
HPLT 2.0	934,236,876	1,152,703	11.71GB
FineWeb2	434,560,077	1,077,501	1.74GB
NSina	94,394,362	486,932	1.87GB
FDC	5,402,768	364,402	142MB
SinMin	104,428,504	313,910	1.85GB
SemiSOLD	1,938,756	107,210	48.5MB

Table 1: Statistics of SINHALA-Corpus1.5B. Any continuous sequence of non-whitespace characters is considered as a token.

4 Sinhala Encoder-only Language Models

Since the introduction of BERT (Devlin et al., 2019), encoder-only transformer-based LMs have dominated most applications in NLP. Despite the rise of large language models (LLMs) such as GPT, encoder-only LMs remain widely used and continue to outperform LLMs in various non-generative NLP tasks, such as text classification (Zampieri et al., 2023; Krugmann and Hartmann, 2024) and sequence labelling (Zaratiana et al., 2024). Therefore, in this research, we focus on building Sinhala encoder-only transformer models.

We train three popular transformer architectures on the corpus we compiled in §3; BERT (Devlin et al., 2019) (Raja), RoBERTa (Liu, 2019) (Koliya)

#	Task	|Train|	|Test|	Splits	Reference	Metric	Domain
				Text Classification			
1	**SA**	7320	1820	💡 ♻	Ranathunga and Liyanage (2021)	Macro F1	News comments
2	**OLD**	7500	2500		Ranasinghe et al. (2024a)	Macro F1	Twitter
3	**NHP**	7870	1970		New	Macro F1	News
				Text Regression			
4	**STS**	5000	100		Kadupitiya et al. (2016)	Spear. Corr.	SICK
				Token Classification			
5	**NER**	4000	1000	♻	Manamini et al. (2016)	Macro F1	News
6	**OTD**	7500	2500		Ranasinghe et al. (2024a)	Macro F1	Twitter

Table 2: Summary of the tasks included in SINHALA-GLUE. The numbers in the |**Train**| and |**Test**| columns are in terms of examples. The **Metric** column shows the primary metric used for evaluation. The **Domain** is based on the source of the texts. ♻ in **Splits** column shows new splits created as the splits are not available. 💡 is a redefined task. NHP task is a new task introduced in this paper.

and `Electra` (Clark et al., 2020) (Mahasen), with the following configurations.

- We select a vocabulary of 64,000 to train the tokeniser. For each model, we train its associated tokeniser from scratch, available through the HuggingFace transformers package.
- We use a maximum sequence length of 512 and a batch size of 64. For the remaining hyperparameters, we used the same given in their English models.
- We train our models on a single NVidia L40 48G GPU. The training took approximately 18 days for each model.

5 Constructing SINHALA-GLUE: Sinhala NLU Benchmark

5.1 SINHALA-GLUE

Table 2 shows the six datasets that are included in SINHALA-GLUE. Table 3 shows examples from each dataset and their corresponding labels. We also show the translations by a native Sinhala speaker in the same table.

5.1.1 Sentiment Analysis (SA)

This dataset released by Ranathunga and Liyanage (2021) focuses on fine-grained sentiment analysis of news comments. The comments were extracted from the online version of Lankadeepa, a local newspaper. All the comments are manually annotated for three classes: *'positive'*, *'negative'* and *'neutral'*.

Ranathunga and Liyanage (2021) originally defined the task as predicting sentiment based on both the news comment and its associated article; however, they used the comment itself as the only input to their machine learning models. After reviewing the dataset, we observed that many comments are highly contextual and closely related to the corresponding news articles. Therefore, we redefined the models to predict sentiment based on both the news comment and its associated article.

5.1.2 Offensive Language Detection (OLD)

This dataset released by Ranasinghe et al. (2024a), also known as SOLD, contains 10,000 Tweets annotated as *'offensive'* or *'not offensive'*. SOLD was part of HASOC 2023 - Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages shared task (Ranasinghe et al., 2024b; Satapara et al., 2023), which was the first ever shared task organised for Sinhala. While several offensive language detection datasets are available for Sinhala, such as Sandaruwan et al. (2019), SOLD is the only publicly available dataset.

5.1.3 News Headline Prediction (NHP)

This is a new dataset constructed for the task of predicting the correct headline for a news article. We construct the dataset using news articles and their headlines from Hettiarachchi et al. (2024), the largest and most recent news corpus released for Sinhala. We created data samples combining news articles with their actual headlines and some incorrect ones. To ensure the incorrect titles are not entirely unrelated to the article, we select them based on a significant word overlap with the original article. Similar tasks have been proposed in NLU benchmarks in other languages (Kakwani et al., 2020).

SA	**news_content:** 149 වැනි පොලිස් විරු සමරු දිනයේ එහි ප්‍රධාන සැමරුම පොලිස්පති එන්.කේ. ඉලංගකෝන් මහතාගේ ප්‍රධානත්වයෙන් බම්බලපිටිය පොලිස් ක්ෂේත්‍ර බළකා මූලස්ථාන පරිශ්‍රයේදී පැවැත්විණ. සෙසු සැමරුම් රට පුරා පැවැත්විණි. ... (The main celebration of the 149th Police Heros Commemoration Day was held at Bambalapitiya Police Field Force Headquarters under the chairmanship of the Inspector General of Police, N. K. Illangakoon. Other celebrations were held across the country. ...) **comment:** සියළුම පොලිස් නිළ දරුවන්ට අපගේ ප්‍රණාමය! (Congratulations to all the police officers!) **sentiment:** POSITIVE
OLD	**tweet:** @USER අපොයි මෙහෙමත් මොඩයෙක්. ජනදිපති අපෙක්ශකයෙක් මීටවඩා බුද්දිමත් විදිහට කතාකලයුතුයි. (@USER what a fool, a presidential candidate should speak intelligently than this.) **label:** OFF
NHP	**news_content** උතුරු, උතුරුමැද, නැගෙනහිර සහ ඌව පළාත්වලට ද හම්බන්තොට දිස්ත්‍රික්කයට ද විටින් විට වැසි ඇති වන බව කාලගුණ විද්‍යා දෙපාර්තමේන්තුව කියයි.'අපරභාගයේදී හෝ සන්ධ්‍යා කාලයේ දි සෙසු ප්‍රදේශවල ද තැනින් තැන ගිගුරුම් සහිත වැසි වර්ධනය වේ'.'කන්කසන්තුරේ සිට ත්‍රිකුණාමලය සහ පොතුවිල් හරහා හම්බන්තොට දක්වා වන මුහුදු ප්‍රදේශවල තැනින් තැන වැසි ඇතිවන අතර දිවයින වටා වන සෙසු මුහුදු ප්‍රදේශ වල අපරභාගයේදී හෝ රාත්‍රී කාලයේදී තැනින් තැන ගිගුරුම් සහිත වැසි ඇති වේ'. ... (The Department of Meteorology says occasional rains will occur in North, North Central, East and Uva provinces and Hambantota district. 'In the afternoon or evening, scattered thunderstorms will develop in other areas too.' 'Scattered rains will occur in the coastal areas from Kankasanthura to Trincomalee and Pottuvil to Hambantota, and there will be scattered thunderstorms in the rest of the coastal areas around the island in the afternoon or at night.' ...) **headline:** ඉදිරි 24 පැයේ කාලගුණය (Weather for the next 24 hours) **is_headline:** 1
STS	**sentence1:** මිනිසුන් තිදෙනෙක් හිම ක්‍රීඩාවේ යෙදෙයි (Three people are playing snow sports) **sentence2:** මිනිසුන් හිම මත ලිස්සා යයි (People are skiing) **similarity:** 0.8
NER	**tokens:** [පිලිපීන, ජනාධිපතිවරණයෙන්, බෙනිග්නෝ, අකීනෝ, ජය, ලබා, ඇති, බවට, වාර්තා, පළ, වේ, .] ([It, is, reported, that, Benigno, Aquino, has, won, the, Philippine, presidential, election, .]) **ner_tags:** [පිලිපීන$^{\text{LOC}}$, ජනාධිපතිවරණයෙන්, බෙනිග්නෝ$^{\text{PER}}$, අකීනෝ$^{\text{PER}}$, ජය, ලබා, ඇති, බවට, වාර්තා, පළ, වේ, .] ([It, is, reported, that, Benigno$^{\text{PER}}$, Aquino$^{\text{PER}}$, has, won, the, Philippine$^{\text{LOC}}$, presidential, election, .])
OTD	**tweet:** @USER අපොයි මෙහෙමත් මොඩයෙක් . ජනදිපති අපෙක්ශකයෙක් මීටවඩා බුද්දිමත් විදිහට කතාකලයුතුයි . (@USER what a fool , a presidential candidate should speak intelligently than this .) **rationales:** @USER අපොයි මෙහෙමත් මොඩයෙක් . ජනදිපති අපෙක්ශකයෙක් මීටවඩා බුද්දිමත් විදිහට කතාකලයුතුයි . (@USER what a fool , a presidential candidate should speak intelligently than this .)

Table 3: Examples from SINHALA-GLUE benchmark. For each task, the last item indicates the label(s) of the given example, and the other items indicate the inputs. English translations by a native speaker are given in brackets.

5.1.4 Semantic Textual Similarity (STS)

The goal of semantic textual similarity is to predict the extent to which two sentences convey the same meaning (Cer et al., 2017) on a scale of 0-1. STS is a popular task in NLU benchmarks such as GLUE (Wang et al., 2018). Kadupitiya et al. (2016) constructed this Sinhala STS dataset with the sentences translated and post-edited from the English SICK dataset (Marelli et al., 2014). This dataset has also been included in the recently released multilingual semantic textual similarity benchmark (MUSTS) (Ranasinghe et al., 2025).

5.1.5 Named Entity Recognition (NER)

This dataset released by Manamini et al. (2016) has named entity recognition annotations for persons (PER), organisations (ORG), and locations (LOC). The sentences are sourced from Sinhala news articles. This is the only Sinhala publicly available dataset for named entity recognition.

5.1.6 Offensive Token Detection (OTD)

This is the second task of the SOLD dataset (Ranasinghe et al., 2024a), where the goal is to predict whether a particular token contributes to the offensiveness of the sentence level if a sentence is offensive. Following this, each token has been annotated as *'offensive'* or *'not offensive'*. This is the only such dataset available for Sinhala.

5.2 Discussion

Machine Translated Datasets - We excluded datasets that were automatically translated from another language. The only exception is the STS dataset (§5.1.4), which originates from translations; however, Kadupitiya et al. (2016) post-edited and re-annotated it. Automatically translated datasets can introduce translation errors and stylistic biases that impact model training and evaluation (Mager et al., 2018), particularly for low-resource languages like Sinhala, where machine translation systems are still evolving (Mahfuz et al., 2025). Consequently, SINHALA-GLUE does not include any automatically translated datasets.

Undocumented Datasets - Several Sinhala datasets have been released on platforms like Kaggle and HuggingFace (Lhoest et al., 2021) without an accompanying published paper. We excluded these datasets from SINHALA-GLUE, as proper documentation is necessary to assess their quality. Only datasets published in peer-reviewed papers were considered.

Code-mixed Datasets - Recently, several Sinhala code-mixed and code-switched datasets, such as Sinhala-CMCS (Rathnayake et al., 2022), have been released. However, we excluded these from the benchmark, as its primary goal is to evaluate the performance of language models on NLP tasks written in the Sinhala script.

Omitted Tasks - We also eliminated two tasks that had datasets satisfying the above requirements; (*i*) News media identification, and (*ii*) News category prediction. Both are text classification tasks that have been included in benchmarks in other languages (Liang et al., 2020). However, Hettiarachchi et al. (2024) demonstrated that language models achieve exceptionally high performance on these tasks for Sinhala, with F1 scores around 0.95. Further analysis of the released Sinhala datasets in these two tasks (Hettiarachchi et al., 2024) revealed that the text contains explicit hints about the news media source and category, making classification trivial for language models. As a result, we excluded these tasks from the benchmark.

Dataset Licenses - We maintain the original licenses assigned by the authors for all datasets included in the SINHALA-GLUE benchmark. All datasets are accessible for research purposes.

Limitations and Comparisons - The SINHALA-GLUE benchmark consists of six NLU tasks, including two token classification tasks, one regression task, and three text classification tasks. While the benchmark encompasses three distinct task types, its scope is limited by the available resources for Sinhala. As a result, certain NLU tasks, such as Question Answering, which are popular in benchmarks in other languages like GLUE (Wang et al., 2018) could not be included. However, we observe similar limitations with other popular benchmarks. For instance, the Bulgarian NLU benchmark (Hardalov et al., 2023) also includes three task types, while the Italian NLU benchmark (Basile et al., 2023) features only two, despite both languages having significantly more resources than Sinhala.

We also acknowledge that some datasets in SINHALA-GLUE can contain bias. For example, in the sentiment analysis task, which is also a highly subjective task, the majority of the instances were annotated by a single annotator (Ranathunga and Liyanage, 2021). While the authors report a high inter-annotator agreement, only a small subset of the dataset has been annotated by both annotators, leaving the rest of the annotations highly biased. However, given the scarcity of publicly available Sinhala datasets, we included it in the benchmark despite these limitations.

Public Leaderboard - Finally, we release SINHALA-GLUE as a public leaderboard following the structure of the existing ones, such as GLUE (Wang et al., 2018). Participants receive access to all training and test examples, but without the gold labels for the test set. They submit a file containing their predictions for each task, which our system then evaluates automatically.

The primary goal of our leaderboard is to provide a standardised framework for comparing model performance on specific Sinhala NLP tasks. This enables researchers and practitioners to assess the current state of the art and identify areas for improvement for Sinhala. However, we caution against drawing broad conclusions about general language understanding solely based on leaderboard performance, whether on our platform or other NLP leaderboards (Ethayarajh and Jurafsky, 2020).

6 Experiments

In this section, we first describe the models we experimented with and then present the evaluation results.

6.1 Models

Baselines - Our baselines include three widely used multilingual encoder-only pre-trained trans-

Task	Input	Output	Loss
SA	[CLS] *news_content* [SEP] *comment*	Positive / Negative / Neutral	Binary Cross Entropy
OLD	[CLS] *Tweet*	Offensive / Not offensive	Binary Cross Entropy
NHP	[CLS] *news_content* [SEP] *headline*	1 / 0	Binary Cross Entropy
STS	[CLS] *sentence1* [SEP] *sentence2*	Similarity (0–5)	Mean Squared Error
NER	[CLS] *news_content*	LOC / ORG / PER / O	Per Token Cross Entropy
OTD	[CLS] *Tweet*	Offensive / not Offensive	Per Token Cross Entropy

Table 4: **Input** format for each task, the special tokens are replaced with the corresponding ones from the baseline model. Expected **Output**, e.g., tag name, class, rating, etc. and the optimisation **Loss** used for training.

former models; `XLM-R` (Conneau et al., 2020), `info-XLM` (Rathnayake et al., 2022) and `RemBERT` (Chung et al., 2021). We did not use the popular `mBERT` as it does not support Sinhala. Additionally, we used `SinBERT` (Dhananjaya et al., 2022), a previously trained transformer model for Sinhala, albeit trained on a relatively small corpus, as a baseline. These models were compared with the three transformer models trained in this paper.

Architecture and Configurations - For all tasks, we introduce a projection layer on top of the representations of the pre-trained language model. For classification tasks (*SA, OLD, NHP*), the output of the *CLS* token maps to the number of classes. For regression (*STS*), we project it to a single continuous value. Finally, for token classification tasks, we apply the classification head on top of each token's representation, which is the first sub-token.

In the following list, we describe the values of the hyperparameters.

- All our models use the AdamW (Loshchilov and Hutter, 2019) optimiser with a weight decay of 1e-8, learning rate of 2e-5, a warmup ratio of 0.06 from the training data and are trained for five epochs with a batch size of 32 (gradient accumulation is applied when needed), and a maximum length of 512 tokens. The values of the hyperparameters (including the number of training epochs) were set to fixed values to ensure consistency across all models.
- All the models were evaluated during training using a development set that consisted of one-fifth of the rows, which were separated from the training set before the start of the training process.
- The best checkpoints were selected on the development set. We use the target metric for each task as a checkpoint selection criterion.
- We trained our models on an NVidia L40 48G GPU. Depending on the dataset size, the experiments took between 20 minutes for the smaller datasets and models and up to 2 hours for the larger datasets.
- All models were trained with half precision (fp16) using the default PyTorch implementation.
- When evaluating the *Token Classification Tasks* if the predicted sequence was shorter than the target one (i.e., not all inputs fit into 512 tokens), we added empty tags ('*O*' or '*not offensive*') until the target length was reached.

The input, output and loss functions used for each task are shown in Table 4.

6.2 Results

Table 5 shows the results for the experimented models fine-tuned on the SINHALA-GLUE tasks. We describe key observations below.

> *Language models introduced in this paper provide the best results in all the tasks in* SINHALA-GLUE

As can be seen in Table 5, `Sinhala-BERT`$_{\texttt{Large}}$ (Raja) trained in this paper provided the best results for all the tasks. The results are closely followed by the other two models trained in this paper as well; `Sinhala-RoBERTa`$_{\texttt{Large}}$ (Koliya) and `Sinhala-Electra`$_{\texttt{Large}}$ (Mahasen).

The models trained in this paper largely outperform the previously trained Sinhala transformer models in all the tasks. Notably, we observe approximately 20% improvements in sentiment analysis (SA) and semantic textual similarity (STS) tasks and approximately 10% improvements in news headline prediction (NHP), named entity recognition (NER) and offensive token detection (OTD).

We attribute this to the fact that we trained the Sinhala LMs in a larger and more diverse corpus compared to `SinBERT`, which resulted in superior LMs.

> *Multilingual LMs provide comparable results for Sinhala NLP tasks.*

As can be seen in the results, all experimented multilingual models consistently provided good

Model Name	Avg. →	SA F1$_{macro}$	OLD F1$_{macro}$	NHP F1$_{macro}$	STS S Corr.	NER F1$_{macro}$	OTD F1$_{macro}$
Multilingual LMs							
`XLM-R`$_{Large}$	79.99	75.27	83.16	77.16	78.28	93.47	72.57
`XLM-R`$_{Base}$	77.53	72.14	81.28	75.19	73.29	92.46	70.79
`info-XLM`$_{Large}$	<u>81.59</u>	<u>77.56</u>	<u>83.89</u>	<u>79.12</u>	<u>79.16</u>	<u>94.03</u>	<u>73.78</u>
`info-XLM`$_{Base}$	79.35	73.64	81.67	76.89	78.89	93.85	71.15
`RemBERT`	80.69	73.45	83.85	78.88	81.06	93.91	72.98
Previous Sinhala LMs							
`SinBERT`$_{Large}$	69.81	61.63	81.12	71.56	59.13	81.08	62.31
`SinBERT`$_{Small}$	66.98	59.11	80.86	69.87	53.55	77.89	60.58
Models from this Paper							
`Sinhala-BERT`$_{Large}$ (Raja)	**82.24**	**79.06**	**84.01**	**80.32**	**82.34**	**94.56**	**75.16**
`Sinhala-RoBERTa`$_{Large}$ (Koliya)	81.75	78.23	83.89	80.16	81.18	94.32	74.67
`Sinhala-Electra`$_{Large}$ (Mahasen)	80.97	78.86	83.78	80.11	81.89	94.15	75.04

Table 5: Model results on the SINHALA-GLUE benchmark. We show the best result for each task in **bold**. We also <u>underline</u> the best result for each task from the multilingual models. The scores for each model are the highest ones achieved by selecting the best model checkpoint on each task's development set. The given scores are percentages following the same notation of previous benchmarks.

results in SINHALA-GLUE. Aligning with the previous research (Hettiarachchi et al., 2024), multilingual models outperformed `SinBERT` models in all the tasks.

Similar to previous research (Devlin et al., 2019), we notice that larger variants of multilingual LMs produce better results in all the tasks.

Construction of SINHALA-GLUE also revealed that `info-XLM` outperforms `XLM-R` in Sinhala NLP tasks, despite the latter's widespread use. We highlight the importance of a well-designed evaluation benchmark in uncovering valuable insights for processing the Sinhala language.

> *Models achieve the best performance on NER, while OTD shows the weakest results.*

Among the various tasks in SINHALA-GLUE, all models achieved the best results for named entity recognition (NER). In contrast, they struggle the most with offensive token detection (OTD), despite both tasks falling under the same category, token classification. The contextual ambiguities associated with offensive tokens can be considered the main reason, making it a more challenging task for the models.

The text classification task, offensive language detection (OLD), achieved the second-best results across all models. Meanwhile, news headline prediction (NHP) and the text regression task, semantic textual similarity (STS), performed comparably across most models. However, sentiment analysis (SA) also proved to be a challenging task, particularly due to its contextual nuances.

Overall, we highlight that SINHALA-GLUE consists of several challenging NLU tasks. We suggest exploring more advanced techniques like contrastive learning (Liang et al., 2024) to tackle these tasks.

7 Conclusions

In this paper, we collected a large Sinhala corpus containing more than 1.5B tokens and trained three popular transformer models on it. We also compiled the first NLU benchmark for Sinhala; SINHALA-GLUE, comprising six tasks. We showed that transformer models trained in this paper, using a large Sinhala corpus, outperform the popular multilingual LMs, and existing Sinhala LMs.

The SINHALA-Corpus1.5B, alongside the three pretrained transformer models, is publicly released[1]. Furthermore, we have open-sourced the datasets in SINHALA-GLUE, incorporating new and redesigned tasks, along with the source code for training and evaluation. Additionally, we released 60 fine-tuned models, one for each task and model combination, all of which are integrated into the HuggingFace Hub. It is the most extensive collection of NLP models for Sinhala. We believe that our paper will foster future advancements in Sinhala natural language processing.

[1]Available at `https://github.com/Sinhala-NLP/Sinhala-GLUE`

In future, we plan to add more tasks for Sinhala with different task types. We also plan to construct a text generation benchmark for Sinhala that could evaluate the performance of large language models.

Limitations

The limitation in SINHALA-GLUE is discussed in Section §5. Additionally, none of the tasks included in SINHALA-GLUE does not belong to a specialised domain such as legal or biomedical. We plan to address this limitation in future work.

As previously discussed, this study focuses on relatively small encoder-only transformer architectures. For future work, we aim to explore alternative modelling approaches and techniques known to enhance efficiency and reduce computational demands, such as few-shot and zero-shot in-context learning, instruction-based evaluation, and multi-task learning.

In this work, we did not investigate whether the datasets contain potential biases, which could contribute to undesirable behaviours in the models trained during our experiments.

Ethical Considerations

All the datasets explored in this paper are publicly available. Furthermore, all the models that we experimented with in this paper are publicly available in HuggingFace (Lhoest et al., 2021). Any new models that we created in this paper, will be made publicly available.

Acknowledgements

We would like to thank the anonymous reviewers for their positive and valuable feedback. We further thank the creators of the datasets used in this paper for making the datasets publicly available for our research.

The experiments in this paper were conducted in UCREL-HEX (Vidler and Rayson, 2024). We would like to thank John Vidler for the continuous support and maintenance of the UCREL-HEX infrastructure, which enabled the efficient execution of our experiments.

The three pre-trained transformer models introduced in this paper are named after three popular tuskers in Sri Lanka, where the majority of Sinhala speakers reside.

1. `Sinhala-BERT`$_{\texttt{Large}}$ (Raja) - Commonly known as Nadungamuwa Raja, is arguably the most popular elephant in Sri Lanka. Raja was considered to be the largest tamed elephant in Asia and was 10.5ft tall. Nadungamuwa Raja died on 7 March 2022, believed to be 68 or 69, following a brief natural illness.

2. `Sinhala-RoBERTa`$_{\texttt{Large}}$ (Koliya) - Koliya was a young tusker in Sri Lanka, known for his unique tusk position. While writing this paper, he was found dead, likely killed by poachers.

3. `Sinhala-Electra`$_{\texttt{Large}}$ (Mahasen) - Mahasen is the oldest tusker with the largest tusks in Sri Lanka.

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.

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Dammi Bandara, Nalin Warnajith, Atsushi Minato, and Satoru Ozawa. 2012. Creation of precise alphabet fonts of early brahmi script from photographic data of ancient sri lankan inscriptions. *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition*, 3(3):33–39.

Valerio Basile, Livio Bioglio, Alessio Bosca, Cristina Bosco, and Viviana Patti. 2023. UINAUIL: A unified benchmark for Italian natural language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 348–356, Toronto, Canada. Association for Computational Linguistics.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

Daniel Cer, Mona Diab, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

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.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

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.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Kushan Hewapathirana, Nisansa de Silva, and CD Athuraliya. 2024. M2ds: Multilingual dataset for multi-document summarisation. In *International Conference on Computational Collective Intelligence*, pages 219–231. Springer.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In

Proceedings of the 37th International Conference on Machine Learning, ICML'20. JMLR.org.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jcs Kadupitiya, Surangika Ranathunga, and Gihan Dias. 2016. Sinhala Short Sentence Similarity Calculation using Corpus-Based and Knowledge-Based Similarity Measures. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 44–53, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junzhe Liang, Haifeng Sun, Zirui Zhuang, Qi Qi, Jingyu Wang, and Jianxin Liao. 2024. Distantly supervised contrastive learning for low-resource scripting language summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5006–5017, Torino, Italia. ELRA and ICCL.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2025. Too late to train, too early to use? a study on necessity and viability of low-resource Bengali LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1183–1200, Abu Dhabi, UAE. Association for Computational Linguistics.

S.A.P.M. Manamini, A.F. Ahamed, R.A.E.C. Rajapakshe, G.H.A. Reemal, S. Jayasena, G.V. Dias, and S. Ranathunga. 2016. Ananya - a Named-Entity-Recognition (NER) system for Sinhala language. In *2016 Moratuwa Engineering Research Conference (MERCon)*, pages 30–35.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages.

Randil Pushpananda, Chamila Liyanage, Ashmari Pramodya, and Ruvan Weerasinghe. 2024. Tamsipara: A tamil–sinhala parallel corpus. In *International Conference on Text, Speech, and Dialogue*, pages 159–170. Springer.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024a. SOLD: Sinhala offensive language dataset. *Language Resources and Evaluation*, pages 1–41.

Tharindu Ranasinghe, Koyel Ghosh, Aditya Shankar Pal, Apurbalal Senapati, Alphaeus Eric Dmonte, Marcos Zampieri, Sandip Modha, and Shrey Satapara. 2024b. Overview of the hasoc subtracks at fire 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 13–15, New York, NY, USA. Association for Computing Machinery.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025. Sinhala encoder-only language models and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Surangika Ranathunga and Isuru Udara Liyanage. 2021. Sentiment Analysis of Sinhala News Comments. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2022. Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowledge and Information Systems*, 64(7):1937–1966.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

H.M.S.T Sandaruwan, S.A.S Lorensuhewa, and M.A.L Kalyani. 2019. Sinhala hate speech detection in social media using text mining and machine learning. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, volume 250, pages 1–8.

Shrey Satapara, Hiren Madhu, Tharindu Ranasinghe, Alphaeus Eric Dmonte, Marcos Zampieri, Pavan Pandya, Nisarg Shah, Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2023. Overview of the hasoc subtrack at fire 2023: Hate-speech identification in sinhala and gujarati. In *FIRE (Working Notes)*, pages 344–350.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Casper S. Shikali and Refuoe Mokhosi. 2020. Enhancing african low-resource languages: Swahili data for language modelling. *Data in Brief*, 31:105951.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

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*.

Menan Velayuthan and Kengatharaiyer Sarveswaran. 2025. Egalitarian language representation in language models: It all begins with tokenizers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5987–5996, Abu Dhabi, UAE. Association for Computational Linguistics.

John Vidler and Paul Rayson. 2024. UCREL - Hex; a shared, hybrid multiprocessor system. `https://github.com/UCREL/hex`.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Yudhanjaya Wijeratne and Nisansa de Silva. 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. *arXiv preprint arXiv:2007.07884*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Yuxian Gu, Kun Zhou, Xuancheng Huang, Wenhao Li, Shuhuai Ren, Jinliang Lu, Chengqiang Xu, Huadong Wang, Guoyang Zeng, Zile Zhou, Jiajun Zhang, Juanzi Li, Minlie Huang, Rui Yan, Xiaodong He, Xiaojun Wan, Xin Zhao, Xu Sun, Yang Liu, Zhiyuan Liu, Xianpei Han, Erhong Yang, Zhifang Sui, and Maosong Sun. 2021. CUGE: A Chinese Language Understanding and Generation Evaluation Benchmark. *arXiv preprint arXiv:2112.13610*.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France. European Language Resources Association.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.