Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7377–7385
Marseille, 20-25 June 2022

BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, Sanath Jayasena
Department of Computer Science and Engineering
University of Moratuwa, Katubedda 10400, Sri Lanka
dhananjayagv.21@uom.lk, piyumalanthony.16@cse.mrt.ac.lk, surangika@cse.mrt.ac.lk, sanath@uom.lk

Abstract

This research provides the first comprehensive analysis of the performance of pre-trained language models for Sinhala text classification. We test on a set of different Sinhala text classification tasks and our analysis shows that out of the pre-trained multilingual models that include Sinhala (XLM-R, LaBSE, and LASER), XLM-R is the best model by far for Sinhala text classification. We also pre-train two RoBERTa-based monolingual Sinhala models, which are far superior to the existing pre-trained language models for Sinhala. We show that when fine-tuned, these pre-trained language models set a very strong baseline for Sinhala text classification and are robust in situations where labeled data is insufficient for fine-tuning. We further provide a set of recommendations for using pre-trained models for Sinhala text classification. We also introduce new annotated datasets useful for future research in Sinhala text classification and publicly release our pre-trained models.

Keywords: Language Models, Monolingual, Multilingual, Text Classification

1. Introduction

Large-scale monolingual pre-trained language models (MonoLMs) such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and their multilingual descendants mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019) (respectively), have shown promising results for high-resource as well as low-resource languages, particularly for text classification (Wu and Dredze, 2019; Aguilar et al., 2020). Following this early success, many empirical studies have been carried out to determine the performance of these models on different settings. However, most of these studies focused on a limited number of languages. Moreover, experimental results show that the performance of the pre-trained Multilingual Language Models (MultiLMs) depends on many factors - such as the amount of language data used during the pre-training, the relatedness of a language to the other languages in the pre-trained model, language characteristics, the typography of the language, and the amount of fine-tuning data used (Wu and Dredze, 2020; Doddapaneni et al., 2021). Thus, the results reported in these studies cannot be generalized across languages. For MonoLMs, a major deciding factor is the amount of monolingual data used in the model pre-training stage (Rust et al., 2020).

As noted by Soria et al. (2018), 'a Natural Language Processing (NLP) system can be measured only in terms of its usefulness for the end-users'. In other words, the usefulness of pre-trained language models on a language depends on their ability to provide acceptable results for the NLP tasks of the considered language, but not by their performance on some other set of languages. Thus, it is imperative that we carry out extensive evaluation of the pre-trained models for the specific languages of interest.

Sinhala is an Indo-Aryan language primarily used by a population of about 20 million, in the small island nation of Sri Lanka. According to Joshi et al. (2020)'s language categorization, Sinhala has been given class 1, meaning an extremely low-resource language. This is not surprising - not only the available language resources, but also the amount of research is scarce for Sinhala (de Silva, 2019). However, Sinhala has been fortunate to get included in pre-trained MultiLMs such as XLM-R, LASER (Artetxe and Schwenk, 2019) LaBSE (Feng et al., 2020), mT5 (Xue et al., 2020) and mBART (Liu et al., 2020)[1]. There also exist monolingual Sinhala pre-trained models[2]. However, they have been pre-trained on relatively small Sinhala corpora. No results of using these monolingual or MultiLMs for Sinhala text classification have been reported so far. Thus, the effectiveness of these models for Sinhala text classification is not known yet.

In this research, we build two RoBERTa based pre-trained language models. Compared to the existing Sinhala pre-trained models, our models are trained with a much larger corpus[3]. Our objective is to identify the best pre-trained model for different Sinhala text classification tasks. Thus, the built Sinhala RoBERTa models are compared against the MultiLMs that include Sinhala; LASER, XLM-R, and LaBSE.

We use 4 different classification tasks, namely, sentiment analysis with a 4-class sentiment dataset (Senevirathne et al., 2020), news category classification with a 5-class dataset (de Silva, 2015b), a 9-class news source classification and a 4-class writing style classification

[1] Sinhala is not included in mBERT

[2] SinBerto; https://huggingface.co/Kalindu/SinBerto and SinhalaBERTo; https://huggingface.co/keshan/SinhalaBERTo

[3] https://github.com/brainsharks-fyp17/sinhala-dataset-creation, corpus at; https://bit.ly/3OBVuoU

task. Dataset for the last two tasks have been prepared by us.

Based on our initial experiment results, we identify XLM-R as the best MultiLM for Sinhala text classification. From the experiments with our MonoLMs and XLM-R, we observe that the XLM-R-large model yields consistently better results than our MonoLMs. We further observe that our MonoLMs perform better than XLM-R-base when the fine-tuning dataset size is small. Based on our results, we provide a set of recommendations, which would be useful for future research on Sinhala text classification. Moreover, we set new baselines for all the selected Sinhala text classification tasks.

We publicly release the trained Sinhala RoBERTa models (which are referred to as SinBERT-large and SinBERT-small, from here onwards) via Hugginface[4], [5]. The annotated datasets for Sinhala news source classification[6], news category classification[7] and writing style classification[8] are also publicly released.

2. Pre-trained Language Models

Pre-trained language models aim at exploiting large unlabeled corpora to learn text representations at scale, such that the trained models can be fine-tuned on relatively smaller, labeled datasets for downstream tasks. LASER and ELMo (Peters et al., 2018) were amongst the initial pre-trained language models based on neural network architectures capable of learning long-term dependencies in sequences, such as GRU (Cho et al., 2014) and Bi-LSTM (Schuster and Paliwal, 1997).

The inception of the Transformer (Vaswani et al., 2017) architecture propelled the creation of state-of-the-art language models. BERT was pre-trained with Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) tasks on the large BookCorpus dataset (Zhu et al., 2015) and English Wikipedia corpus. It became a basis for many other language models that followed up. RoBERTa, which stands for Robustly Optimized BERT Pre-training Approach, is a Transformer architecture similar to BERT. The first RoBERTa model trained for the English language introduced modifications over the pre-training process used in BERT. These include the use of larger batch sizes, removal of the NSP pre-training objective, using longer sequences for pre-training, training the model for a longer period and using a dynamic masking pattern for the MLM task.

[4]SinBERT-small-https://huggingface.co/NLPC-UOM/SinBERT-small

[5]SinBERT-large-https://huggingface.co/NLPC-UOM/SinBERT-large

[6]https://huggingface.co/datasets/NLPC-UOM/Sinhala-News-Source-classification

[7]https://huggingface.co/datasets/NLPC-UOM/Sinhala-News-Category-classification

[8]https://huggingface.co/datasets/NLPC-UOM/Writing-style-classification

Multilingual BERT (mBERT) was released as the multilingual variant of BERT pre-trained on corpora from 104 languages. LaBSE is another MultiLM, which uses a dual-encoder architecture based on BERT and supporting 109 languages. It has been trained on MLM and Translation Language Model (TLM) objectives. XLM-R is a MultiLM pre-trained on CommonCrawl (Wenzek et al., 2019) based on XLM (Lample and Conneau, 2019), and supports 100 languages. It uses MLM as its pre-training task.

While models such as mBERT and XLM-R are encoder-only models, T5 (Roberts et al., 2020) and BART (Lewis et al., 2019) (and their multilingual variants mT5 and mBART) contain an encoder and a decoder, and are ideal for sequence-to-sequence tasks such as text summarization. However, their usage in text classification tasks is comparatively less.

3. Related Work

3.1. Text Classification with Pre-trained Models

Following the success of pre-trained models for English, similar models have been built for some other languages. Some examples are, FlauBERT (French) (Le et al., 2019), FinBERT (Finnish) (Virtanen et al., 2019), AraBERT (Arabic) (Antoun et al., 2020), PhoBERT (Vietnamese) (Nguyen and Nguyen, 2020) and AfriBERT (Afrikaans) (Ralethe, 2020). Each model has been able to set a new state-of-the-art for a variety of NLP tasks for the corresponding language. Some have compared the MonoLMs they built with the MultiLMs (Nguyen and Nguyen, 2020; Le et al., 2019; Virtanen et al., 2019; Ács et al., 2021). However, it cannot be concluded that the MonoLMs are better than MultiLMs, and vice-versa, across different tasks and languages. Wu and Dredze (2020) attributed this discrepancy solely to the amount of data used to pre-train the models. However, Rust et al. (2020)'s findings suggest that the pre-trained tokenizer also plays an important role in downstream task performance, as well as the selected tasks.

In addition to the above research that compared monolingual and MultiLMs, there is a plethora of research that analysed various aspects of pre-trained MultiLMs across various NLP tasks and languages. Aguilar et al. (2020) and Lauscher et al. (2020) showed that the performance of the multilingual pre-trained models is not consistent across the NLP tasks. According to Aguilar et al. (2020), these models are better at syntactic analysis as opposed to semantic analysis. Groenwold et al. (2020) and Lauscher et al. (2020) showed that the performance of a pre-trained model on a given language is heavily influenced by the language family. In other words, more related languages are included in the model is beneficial for a language. As a result, pre-trained models have been shown to perform better for Indo-European languages (Hu et al., 2020). Some others experimented on different conditions such as zero

shot performance on languages that are included in the pre-trained model (Hu et al., 2020; Wu and Dredze, 2019; Ebrahimi et al., 2021; Litschko et al., 2021), and performance on languages not included in the pre-trained models (Ebrahimi and Kann, 2021).

Although pre-trained MultiLMs such as mBERT and XLM-R are a very attractive option for low-resource language computing, they have bounded capacity with respect to the number of languages that can be included in the model. This is commonly known as the "curse of multilinguality" (Conneau et al., 2019). Moreover, low resource languages are mostly underrepresented in MultiLMs (i.e. the pre-trained models include comparatively low amounts of training data from these languages), which makes these models to under-perform for those languages compared to high resource languages included in MultiLMs (Wu and Dredze, 2020). The alternative is to train MultiLMs only for a set of related languages. IndicBERT (Kakwani et al., 2020)[9] is a very good example for this. When the average result for a particular task across the indic languages is considered, IndicBERT outperforms both mBERT and XLM-R by a substantial margin in tasks such as question answering and cross-lingual sentence retrieval.

3.2. Sinhala Text Classification

Being a fusional language and having rich linguistic features, the Sinhala language inherits a certain complexity of language understanding added to its scarcity of resources. Research in Sinhala text classification has been mainly limited to traditional approaches. Experiments with Machine Learning methods such as Support Vector Machines (SVM) were carried out by (Gallege, 2010). Furthermore, approaches such as rule-based systems (Lakmali and Haddela, 2017), a stop word extraction method for text classification using TF-IDF (Gunasekara and Haddela, 2018), Feed-forward Neural network based system (Medagoda, 2017) and a Word2Vec based approach[10] have also been followed. Chathuranga et al. (2019) proposed a method for Sinhala text classification based on a lexicon. Ranathunga and Liyanage (2021) are the first to experiment with Deep Learning techniques such as LSTM networks as well as Convolutional Neural Networks (CNN) based methods for Sinhala sentiment classification. Demotte et al. (2020) also proposed a LSTM-based system for Sinhala text classification based on S-LSTMs (Zhang et al., 2018). More recently, Senevirathne et al. (2020) empirically analysed RNN, Bi-LSTM and Capsule Networks for Sinhala news text sentiment classification. SinBERTo and Sinhala-RoBERTa are two separately pre-trained RoBERTa based MonoLMs for Sinhala, which have been released recently. They do not have related work published, nor have been used in text classification, to the best of our knowledge. Although encoder-based pre-trained models have not been used for Sinhala, mBART has shown exceptionally good results for Machine Translation that involves Sinhala (Thillainathan et al., 2021).

[9] https://indicnlp.ai4bharat.org/indic-bert/

[10] http://bit.ly/2QKI9Np

4. SinBERT Model

4.1. Pre-training and Fine-tuning Setup

RoBERTa has shown improved results over other competitive models for the GLUE benchmark (Wang et al., 2018), specifically for classification tasks. Hence, we build our Sinhala MonoLMs based on RoBERTa. We use Huggingface's[11] Transformers libraries in Pytorch (Paszke et al., 2019) to pre-train our RoBERTa models[12]. We use AdamW (Loshchilov and Hutter, 2017) as the optimizer with a learning rate of 1e-4, a batch size of 16 and a maximum of 2 training epochs to pre-train the models. We introduce two variants of our model; SinBERT-small containing 6 hidden layers and SinBERT-large containing 12 hidden layers. Parameters of the two models are shown in Table 2.

Fine-tuning hyper-parameters are given in Table 3. We used the standard fine-tuning process, where the [CLS] token output from the pre-trained model's encoder was fed to a feed-forward neural network based classifier. For Sinhala monolingual models, we use Huggingface's default classifier for RoBERTa models which consists of a linear layer, a dropout layer preceded by the pooled output from the model encoder layer. A linear layer preceded by a dropout layer was used as the classifier head for LASER and LaBSE. For XLM-R-large, we use a batch size of 8 due to hardware constraints.

We report the macro-averaged F1-score over 5 different randomly-initialized runs for each experiment using 4:1 train/test splits of the datasets. For LaBSE and LASER we use only 3 randomly-initialized runs as their performance is well below to that of XLM-R and the monolingual models. All the pre-training and fine-tuning were conducted on a single Nvidia Quadro RTX 6000 (24GB) GPU.

4.2. Sinhala Corpus used for SinBERT pre-training

SinBERT models are pre-trained using "sin-cc-15M" corpus[13]. At present, it is the largest Sinhala monolingual corpus available to the best of our knowledge. The dataset comprises of 15.7 million sentences extracted from 3 sources: CC-100, OSCAR and raw text data from Sinhala news web sites. CC-100 dataset contains 3.7GB of data for Sinhala and OSCAR contains 802MB of Sinhala text including duplicated text. The raw news data extracted from Sinhala news sites is 413MB in size. The final sin-cc-15M dataset has been cleaned of other language words/characters and invalid

[11] https://huggingface.co/

[12] We publicly release the pre-training and fine-tuning codes on https://github.com/nlpcuom/Sinhala-text-classification

[13] anonymous

characters. Cleaned dataset statistics are shown in Table 1.

Number of words	192.6M
Number of unique words	2.7M
Number of sentences	15.7M
Average number of words/sentence	12.2

Table 1: Statistics of the pre-training corpus

	SinBERT-small	SinBERT-large
Hidden layers	6	12
Attention heads	6	12
Max. Position embeddings	514	514
Vocabulary size	30000	52000
Number of Parameters	66.5M	125.9M

Table 2: Parameters of two SinBERT models

5. Experiments

5.1. Model Selection

We compare the trained RoBERTa models with three MultiLMs: XLM-R-base and large, LaBSE[14], and LASER[15]. For other Sinhala MonoLMs, we take two RoBERTa based models publicly available in Huggingface; SinBERTo and SinhalaBERTo. Both have a vocabulary size of 52 000 and a similar model architecture (6 hidden layers, 12 attention heads, max. position embedding size of 514). However, SinBerto has been trained on a small news corpus while SinhalaBERTo has been trained on a much larger deduplicated Sinhala OSCAR dataset. There are two other Sinhala MonoLMs available in Huggingface (sinhala-Roberta-Oscar[16] and sinhala-roberta-mc4[17]), however, their vocabulary sizes are smaller.

5.2. Fine-tuning Tasks

We use four sentence/document level classification tasks. For the first two tasks given below, annotated data was already available. For the other two tasks, we prepared the annotated data from the raw corpora.

5.2.1. Sentiment Analysis

We use the sentiment dataset published by Senevirathne et al. (2020) for the sentiment classification task. This dataset consists of user comments published in response to online news articles. Each user comment is labeled using four classes (*positive, negative, neutral, conflict*). Thus this can be considered as a document classification task. This is an extension to the dataset introduced by Ranathunga and Liyanage (2021), and carries a Cohen's Kappa value of 0.65. Senevirathne et al. (2020) reported a baseline for this task using RNNs and capsule networks.

5.2.2. News Category Classification

The news category dataset[18] contains sentences extracted from 5 different categories of news (*Business, Political, Entertainment, Science and Technology, Sports*) with 1019 maximum number of sentences and 438 minimum sentences for a class (de Silva, 2015b). Thus this is a sentence classification task. The publicly available version of the news category dataset has not been processed. Hence, we pre-process the dataset and remove sentences that contain English only words and sentences having a length less than 3 words (e.g.- Names of places, celebrities). de Silva (2015b) reported an accuracy score result as a baseline for this task using an approach based on SAFS3 algorithm (de Silva, 2015a).

5.2.3. Writing Style Classification

We extracted text from Upeksha et al. (2015)'s large Sinhala corpus[19], which contains text spanning across a set of genres. For writing style classification, we select text belonging to 4 categories (*News, Academic, Blog, Creative*). This is a document classification task. We process the extracted text by deduplicating, removing English only text and very long text (length larger than 3500 characters). Since the dataset contains long text, we use truncation to fit them into the models. No evaluation has been presented for this dataset.

5.2.4. News Source Classification

This is an annotated dataset newly compiled by us. The news source dataset comprises news headlines in Sinhala, scraped from 9 different Sinhala news web sites (Sri Lanka Army[20], Dinamina[21], GossipLanka[22], Hiru[23], ITN[24], Lankapuwath[25], NewsLK[26], Newsfirst[27], World Socialist Web Site-Sinhala[28]) on the Internet. We reduce the amount of data in the original web-scraped news-source dataset (Sachintha et al., 2021) in order to handle the class imbalance. We also remove one news source (Sinhala Wikipedia) from the originally scraped dataset as it mostly contains invalid characters, numbers and single word sentences. This

[14]https://tfhub.dev/google/LaBSE/2

[15]https://github.com/facebookresearch/LASER

[16]https://huggingface.co/keshan/sinhala-roberta-oscar

[17]https://huggingface.co/keshan/sinhala-roberta-mc4

[18]https://osf.io/tdb84/

[19]https://osf.io/a5quv/files/; publicly available files only contain a portion of the corpora described in their paper

[20]https://www.army.lk/

[21]http://www.dinamina.lk/

[22]https://www.gossiplankanews.com/

[23]https://www.hirunews.lk/

[24]https://www.itnnews.lk/

[25]http://sinhala.lankapuvath.lk/

[26]https://www.news.lk/

[27]https://www.newsfirst.lk/

[28]https://www.wsws.org/si

is a sentence classification task (since we classify the news headings).

Parameter	SinBERT	XLM-R	Other
Starting learning rate	1e-5	5e-6	5e-5, 1e-5
Batch size	16	16, 8	16
No. of epochs	10	5	5
Optimizer	AdamW	AdamW	AdamW

Table 3: Hyperparameters for model fine-tuning

Dataset	No.of data points	No. of classes	Average text length
Sentiment	15059	4	21.66
News sources	24093	9	8.42
News categories	3327	5	23.49
Writing style	12514	4	181.97

Table 4: Statistics of the Fine-tuning datasets used

Dataset	Maximum data points	Minimum data points
Sentiment	7665	1911
News sources	3109	1541
News categories	1019	438
Writing style	4463	2111

Table 5: Statistics of the Fine-tuning datasets used-max/min data points in a class

5.3. Evaluation

Table 6 reports the results for each of our tasks performed using the selected models. We also report the current baseline results for each of the tasks, whenever available. Note that the baseline results for sentiment analysis has been reported with weighted-F1. For a meaningful comparison, we report the same results in the metric used in the baseline paper as well. Since the largest selected model (XLM-R-large) demands high levels of GPU resources to run on, we limit XLM-R-large to the experiments reported in Table 6.

Results of LaBSE and LASER are consistently lower than both our MonoLMs, as well as the XLM-R models across all the tasks. In fact, LaBSE has a very poor performance across all the tasks. Thus, we can safely advise against using these models for Sinhala text classification. XLM-R-large outperforms the base version in all the tasks, which is not surprising. However, this margin is small in tasks such as sentiment analysis and news category classification.

The SinBERT models outperform the existing Sinhala pre-trained models, thus establishing our models as the best monolingual pre-trained models for Sinhala text classification. Interestingly, our large model has only a very small gain against the small model. We believe this is due to the small size of the Sinhala corpus used to pre-train the models- the dataset is not sufficient to properly train the large model. Considering the low performance gains and the time and memory complexity of fine-tuning the SinBERT-large model, we advise the use of the small model in future Sinhala text classification tasks.

It can be seen that the XLM-R-large model outperforms both of our SinBERT models. Thus, if the hardware requirements (see Section 4.1) can be satisfied, the best model choice for Sinhala text classification is the XLM-R-large model. However, in a constrained hardware setting, either the XLM-R-base model or the SinBERT-small model can be used. Specifically, XLM-R-base model outperforms the SinBERT-small for all the tasks except the news source categorisation task. We believe this is because the raw news source dataset was included in the SinBERT model training. This is also an important finding. Even if the annotated data amount is small, if the corresponding raw corpus can be included while model pre-training, a result increase can be expected.

In the XLM-R models, Sinhala data attributes to only $\sim$0.15% of the total pre-trained corpora. Moreover, Sinhala has its own script and characteristics. Compared to this low representation and the uniqueness of the language, XLM-R performance on Sinhala is impressive. Sinhala is an Indo-Aryan language and the model contains a relatively higher proportion of data from related languages such as Hindi, and an even higher proportion of distantly related Indo-European languages. This might have contributed to the high performance gains for Sinhala.

Figures 1 - 4 depict the macro-F1 score for XLM-R-base and SinBERT models with varying dataset sizes. We vary the dataset sizes as 100, 500, 1000, 10000 and total dataset size (for the news type categorization experiment, the experiment with dataset size of 10000 is skipped since its total dataset size is below 10000). All the graphs show that for smaller dataset sizes, XLM-R-base model lags behind SinBERT models but catches up quickly as the dataset size increases. Thus, if the annotated dataset is extremely small, using the SinBERT-small model would be more fruitful.

Even the XLM-R-base model and the SinBERT-small model outperform the current baselines for sentiment analysis. Finally, text classification with the XLM-R-large results establish a new (strong) baseline for each of the considered tasks.

Out of the four contrasting classification tasks and datasets that were used, sentiment analysis and news source classification task yield the lowest F1-scores, thus they can be considered as the most difficult tasks. News source prediction is a difficult task for humans

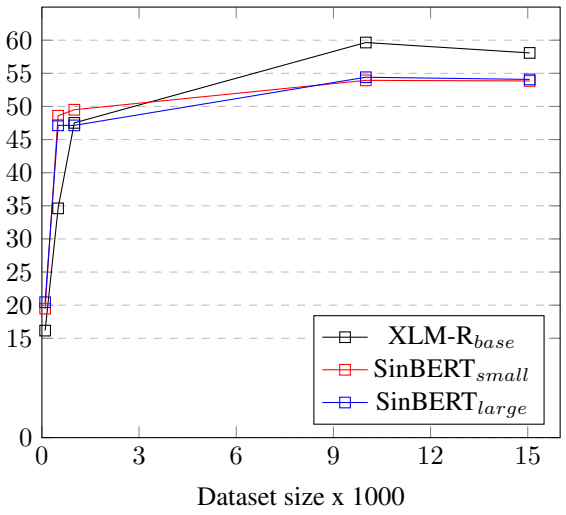

Figure 1: Results in macro-F1 score for varying dataset sizes in sentiment classification task with SinBERT models and XLM-R-base

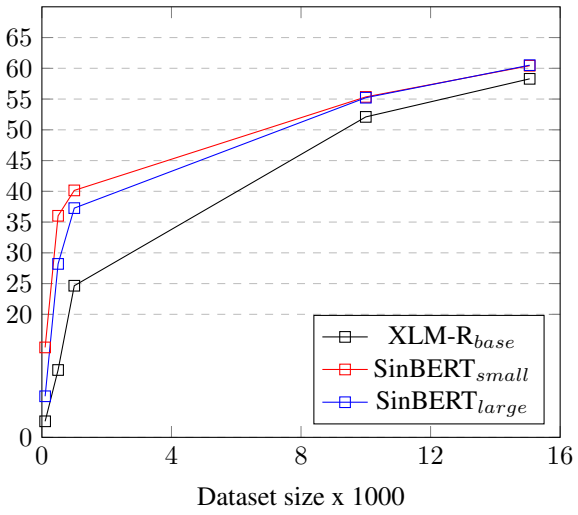

Figure 2: Results in macro-F1 score for varying dataset sizes in news source classification task with SinBERT models and XLM-R-base

as well unless the news headlines carry distinguished styles of writing or keywords in them. In our dataset, Army news website headlines are comparatively shorter in length and contains a small set of frequently used words, which makes it easier to be identified by the model. The sentiment analysis dataset contains one under-represented class label *conflict*, which makes it more challenging for the model to differentiate between the sentiment classes. In the news categories and writing style datasets, the sentences/documents in both datasets contain distinct sets of words or keywords, which makes it easier for the model to predict the classes.

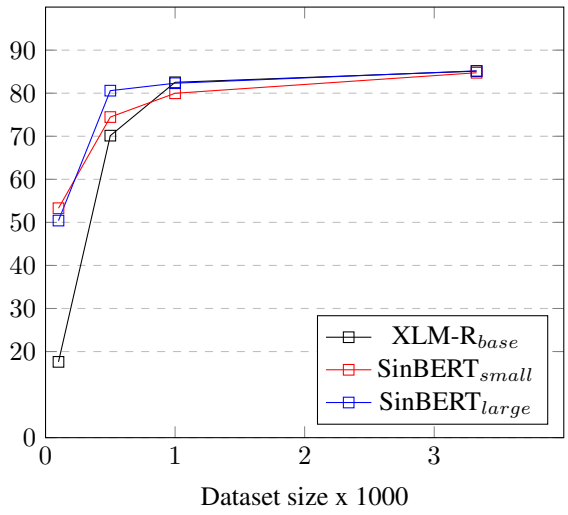

Figure 3: Results in macro-F1 score for varying dataset sizes in news category classification task with SinBERT models and XLM-R-base

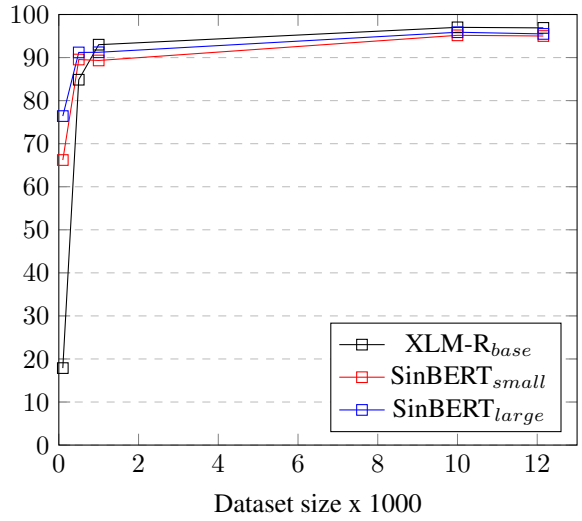

Figure 4: Results in macro-F1 score for varying dataset sizes in writing style classification task with SinBERT models and XLM-R-base

6. Conclusion

Although Sinhala has been included in several multilingual pre-trained language models and there exist several monolingual Sinhala pre-trained models, no empirical analysis has been conducted on their performance with respect to NLP tasks. This paper took the first step in this direction, by providing a comprehensive analysis of these models for Sinhala text classification. We also built two Sinhala pre-trained models, which have been publicly released along with the fine-tuned models. Based on the results, we provided a set of recommendations for future research that plans to use the pre-trained models for Sinhala text classification. We also showed that the XLM-R-large model sets a very strong baseline for Sinhala text classification. As an additional

Model	Sentiment	News sources	News categories	Writing style
Baseline	$59.42_{w.F1}$	-	-	-
LaBSE	20.63	11.85	24.09	-
LASER	54.07	28.84	48.54	87.06
XLM-R_{base}	58.08	58.29	85.12	96.89
XLM-R_{large}	**60.45** ($68.1_{w.F1}$)	**61.84**	**89.54**	**98.41**
SinBERTo	50.83	57.22	78.07	93.84
SinhalaBERTo	49.71	57.34	82.73	94.10
SinBERT$_{small}$	53.85	60.42	84.75	95.00
SinBERT$_{large}$	54.08	60.51	85.19	95.49

Table 6: macro-F1 scores for the selected models on 4 classification tasks.

contribution, we release annotated datasets for Sinhala news source classification and other modified datasets (news category classification, writing style classification) that we use in our experiments. Additionally, we publicly release pre-training and fine-tuning codes. In the future, we plan to improve SinBERT with additional pre-training data and to test on more downstream tasks.

7. Acknowledgement

Vinura Dhananjaya was funded by a Senate Research Committee (SRC) grant of University of Moratuwa, Sri Lanka. We also acknowledge the Prof. V.K. Samaranayake research grant provided by LK Domain Registry for funding the conference participation.

8. Bibliographical References

Ács, J., Lévai, D., and Kornai, A. (2021). Evaluating transferability of bert models on uralic languages. *arXiv preprint arXiv:2109.06327*.

Aguilar, G., Kar, S., and Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France, May. European Language Resources Association.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Chathuranga, P., Lorensuhewa, S., and Kalyani, M. (2019). Sinhala sentiment analysis using corpus based sentiment lexicon. In *2019 19th international conference on advances in ICT for emerging regions (ICTer)*, volume 250, pages 1–7. IEEE.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

de Silva, N. H. N. D. (2015a). Safs3 algorithm: Frequency statistic and semantic similarity based semantic classification use case. In *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 77–83.

de Silva, N. (2015b). Sinhala text classification: Observations from the perspective of a resource poor language. 06.

de Silva, N. (2019). Survey on publicly available sinhala natural language processing tools and research. *CoRR*, abs/1906.02358.

Demotte, P., Senevirathne, L., Karunanayake, B., Munasinghe, U., and Ranathunga, S. (2020). Sentiment analysis of sinhala news comments using sentence-state lstm networks. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 283–288. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Doddapaneni, S., Ramesh, G., Kunchukuttan, A., Kumar, P., and Khapra, M. M. (2021). A primer on pretrained multilingual language models. *CoRR*, abs/2107.00676.

Ebrahimi, A. and Kann, K. (2021). How to adapt your pretrained multilingual model to 1600 languages. 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 4555–4567, Online, August. Association for Computational Linguistics.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., Ortega, J., Ramos, R., Rios, A., Vladimir, I., Giménez-Lugo, G. A., Mager, E., Neubig, G., Palmer, A., Solano, R. A. C., Vu, N. T., and Kann, K. (2021). Americasnli: Evaluating zero-

shot natural language understanding of pretrained multilingual models in truly low-resource languages. *CoRR*, abs/2104.08726.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Gallege, S. (2010). Analysis of sinhala using natural language processing techniques. *University of Wisconsin*.

Groenwold, S., Honnavalli, S., Ou, L., Parekh, A., Levy, S., Mirza, D., and Wang, W. Y. (2020). Evaluating the role of language typology in transformer-based multilingual text classification. *CoRR*, abs/2004.13939.

Gunasekara, S. and Haddela, P. S. (2018). Context aware stopwords for sinhala text classification. In *2018 National Information Technology Conference (NITC)*, pages 1–6. IEEE.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Lakmali, K. and Haddela, P. S. (2017). Effectiveness of rule-based classifiers in sinhala text categorization. In *2017 National Information Technology Conference (NITC)*, pages 153–158. IEEE.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November. Association for Computational Linguistics.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Litschko, R., Vulic, I., Ponzetto, S. P., and Glavas, G. (2021). Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. *CoRR*, abs/2101.08370.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Medagoda, N. P. K. (2017). *Framework for sentiment classification for morphologically rich languages: A case study for Sinhala*. Ph.D. thesis, Auckland University of Technology.

Nguyen, D. Q. and Nguyen, A. T. (2020). Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Ralethe, S. (2020). Adaptation of deep bidirectional transformers for afrikaans language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2475–2478.

Ranathunga, S. and Liyanage, I. U. (2021). Sentiment analysis of sinhala news comments. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–23.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2020). How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.

Sachintha, D., Piyarathna, L., Rajitha, C., and Ranathunga, S. (2021). Exploiting parallel corpora

to improve multilingual embedding based document and sentence alignment. *CoRR*, abs/2106.06766.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Senevirathne, L., Demotte, P., Karunanayake, B., Munasinghe, U., and Ranathunga, S. (2020). Sentiment analysis for sinhala language using deep learning techniques. *arXiv preprint arXiv:2011.07280*.

Soria, C., Quochi, V., and Russo, I. (2018). The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Thillainathan, S., Ranathunga, S., and Jayasena, S. (2021). Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437. IEEE.

Upeksha, D., Wijayarathna, C., Siriwardena, M., Lasandun, L., Wimalasuriya, C., de Silva, N., and Dias, G. (2015). Implementing a corpus for sinhala language. 01.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (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, November. Association for Computational Linguistics.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, November. Association for Computational Linguistics.

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

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhang, Y., Liu, Q., and Song, L. (2018). Sentence-state lstm for text representation. *arXiv preprint arXiv:1805.02474*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.