Instructions for PACLIC 2023 Proceedings

Anonymous PACLIC submission

Abstract

Sentiment analysis has witnessed significant advancements with the emergence of deep learning models such as transformer models. Transformer models adopt the mechanism of self-attention and have achieved state-of-the-art performance across various natural language processing (NLP) tasks, including sentiment analysis. However, limited studies are exploring the application of these recent advancements in sentiment analysis of Sinhala text. This study addresses this research gap by employing transformer models such as BERT, DistilBERT, RoBERTa, and XLM-RoBERTa (XLM-R) for document-level sentiment analysis of Sinhala News comments. This study revealed that the XLM-R-large model outperformed the other three models and the traditional machine learning techniques used in previous studies for the Sinhala language. The XLM-R-large model achieved an accuracy of 63.44% and a macro-F1 score of 67.19% for sentiment analysis with four classes and an accuracy of 71.75% and a macro-F1 score of 68.08% for three classes.

1 Introduction

Sentiment analysis is a fundamental task in NLP which aims to analyze and understand the sentiment expressed in textual data. While sentiment analysis has been extensively studied for major languages such as English, research on sentiment analysis in low-resource languages is relatively limited.

Sinhala, a morphologically rich Indo-Aryan language, serves as the native language of the Sinhalese people, constituting a significant portion of the population in Sri Lanka with an estimated count of 20 million speakers. However, despite its large speaker base, Sinhala is considered a low-resource language in the context of NLP research due to the scarcity of available linguistic resources for analysis and processing.

Sentiment analysis has experienced significant progress with the advent of large-scale pre-trained language models. These models have demonstrated promising results in text classification tasks for high-resource and low-resource languages. Transformer models have revolutionized NLP tasks by leveraging attention mechanisms and self-attention layers, allowing them to capture intricate linguistic patterns and dependencies. Notably, transformer-based models such as BERT, RoBERTa, and XLM-R have shown remarkable performance across various languages, making them promising candidates for sentiment analysis in Sinhala.

One of the primary advantages of employing transformer models for sentiment analysis in Sinhala is their ability to handle the language's morphological richness and syntactic complexities. Sinhala exhibits complicated morphological variations and context-dependent sentiment expressions, which transformer models can effectively capture.

However, applying transformer models to Sinhala sentiment analysis also poses specific challenges. One major challenge is the scarcity of annotated sentiment datasets for fine-tuning transformer models. There exists a sentiment dataset of 15,059 Sinhala News comments, annotated with four classes: Positive, Negative, Neutral, and Conflict. However, the limited size of this dataset hinders the ability of transformer models to achieve optimal performance.

To address this limitation, we expanded the existing Sinhala News comments dataset by adding 5,000 annotated comments to the dataset. While the dataset size may still be considered limited, this extension introduced more diverse examples and enabled some level of expansion for training and evaluation purposes.

In this research, we conducted two sentiment analysis experiments considering four sentiment classes and three sentiment classes respectively. The goal was to evaluate the performance of monolingual models such as BERT, DistilBERT, and RoBERTa as well as multilingual models such as XLM-R-base and XLM-R-large models in sentiment analysis for the Sinhala language. We investigated their capabilities in effectively capturing sentiment information, accommodating the morphological variations of the language, and addressing the limited availability of labeled data. These research outcomes will contribute valuable insights to the field of sentiment analysis in Sinhala and will provide a foundation for future studies and applications.

2 Related Work

Recent developments in deep learning techniques have made it possible to achieve better results in the domain of NLP. Deep learning techniques do not use language-dependent features. Therefore, deep learning techniques have outperformed traditional statistical machine learning techniques [18]. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) were the most popular deep learning techniques used in the NLP domain until Long Short-Term Memory (LSTM) and Transformer models were introduced. Kim proposed a method using CNN with hyperparameter tuning for sentiment analysis [19], and it was shown that a simple CNN with one layer of convolution and little hyperparameter tuning performs remarkably well. LSTM encoders were experimented for sentiment analysis by Yang et al. [20] and bi-directional LSTM by Xu et al. [21]. Both studies showed improved results compared to previous studies, which used deep learning techniques such as CNN and RNN. An attention-based Bi-LSTM with a convolutional layer scheme called AC-BiLSTM was proposed by Liu et al. [22] for sentiment analysis. Word2Vec, which is one of the most popular word-embedding models was introduced in 2013 by Mikolov et al. [23]. Word2Vec improved the efficiency of the training procedure and enhanced the training speed and accuracy. An improved version of the Word2Vec model called GloVe was introduced by Pennington et al. in 2014. GloVe outperformed other models on word analogy, word similarity, and named entity recognition tasks. Transformer models were introduced by Vaswani et al. in 2017 [8]. Transformers could train significantly faster than architectures based on recurrent or convolutional layers. Xu et al. [24] carried out aspect-based sentiment analysis using the BERT model, producing state-of-the-art performance for sentiment analysis. Liao et al. [25] used RoBERTa, an improved version of BERT, to carry out aspect-category sentiment analysis and it outperformed other models for comparison in aspect-category sentiment analysis.

Since Sinhala is a low-resource language, research done on the Sinhala language is very limited. The first sentiment analysis for the Sinhala language was carried out by Medagoda et al. [26] by constructing a sentiment lexicon for Sinhala with the aid of the SentiWordNet 3.0, an English sentiment lexicon. It achieved a maximum accuracy of 60% in Naïve Bayes (NB) classification. The first sentiment analysis for the Sinhala language using an artificial neural network was conducted by Medagoda et al. [27] using a simple feed-forward neural network and part of speech tags as a feature. This model achieved an accuracy of 55% and an F value of 0.51. Chathuranga et al. [28] used a rule-based technique for binary sentiment classification of Sinhala News comments. In this study, Chathuranga et al. generated a Sinhala sentiment lexicon in a semi-automated way and used it for sentiment classification of Sinhala News comments. NB, Support Vector Machines (SVM), and decision trees were used in this study and obtained accuracies between 65% - 70%. Chathuranga et al. obtained the best accuracy of 69.23% for the NB model. Ranathunga and Liyanage [6] conducted sentiment analysis for Sinhala News comments with deep learning techniques such as LSTM and CNN+SVM. Also, this study experimented with Word2Vec and fastText word embeddings for Sinhala [6].

Further, statistical machine learning algorithms such as NB, logistic regression, decision trees, random forests, and SVM were experimented by training them with the same features and conducting a sentiment analysis for Sinhala News comments. This research was carried out to study the use of various models with respect to the dimensionality of the embeddings and the effect of punctuation marks [6]. Demotte et al. [29] used an approach based on the S-LSTM model for sentiment analysis of Sinhala News comments. The same dataset used by Ranathunga and Liyanage [6] was used in this study, and it was found that S-LSTM outperforms the traditional LSTM used in the study conducted by Ranathunga and Liyanage [6]. Senevirathne et al. [30] conducted comprehensive research on the use of RNN, LSTM, and Bi-LSTM models as well as more recent models such as hierarchical attention hybrid neural networks and capsule networks for sentiment analysis. This study released a dataset of 15059 Sinhala News comments, annotated with four classes (Positive, Negative, Neutral, and Conflict) and a corpus of 9.48 million tokens [30]. Dhananjaya et al. [111] conducted experiments to explore the performance of transformer models in various linguistic tasks, including sentiment analysis, for the Sinhala language. Their study evaluated LASER, LaBSE, XLM-R-large, XLM-R-base, and three RoBERTa-based models pre-trained specifically for Sinhala: SinBERT, SinBERTo, and SinhalaBERTo.

3 Models

In this study, we used the following transformer models to carry out sentiment analysis for the Sinhala language,

- BERT (Bidirectional Encoder Representations from Transformers)
- DistilBERT (Distilled version of BERT)
- RoBERTa (Robustly Optimized BERT Pretraining Approach)
- XLM-R (Cross-lingual Language Model – RoBERTa)
  - XLM-R-base
  - XLM-R-large

BERT, which stands for Bi-directional Encoder Representations from Transformers, is a bidirectional transformer model pre-trained on Toronto Book Corpus and Wikipedia. BERT was developed by Google, and it was the state-of-art language model for NLP tasks at the time it was released [4].

DistilBERT is a lighter and faster version of the BERT model, and it was developed by Huggingface. DistilBERT has the same general architecture as BERT, but the size is 40% less than that of BERT and retains 97% of the language understanding capabilities of BERT. Also, DistilBERT is 60% faster than BERT, which is another benefit of this model [13].

RoBERTa stands for Robustly Optimized BERT Pre-training Approach. It is an improved version of the BERT model. RoBERTa has the same architecture as the BERT model but is trained with more data and has better parameter settings [15].

XLM-R is a multilingual model pre-trained on filtered Common Crawl data containing more than 100 languages, including Sinhala. This model was developed and released by Facebook AI in 2019 [16]. XLM-R model outperformed the multilingual BERT (mBERT) and achieved state-of-the-art results on multiple cross-lingual benchmarks [16]. This model can be directly fine-tuned for a downstream task without pre-training on a Sinhala corpus, as this model is already pre-trained on Sinhala. XLM-R consists of two variants: XLM-R-base and XLM-R-large. XLM-R-base is the base version with fewer parameters.

4 Datasets

This study required two datasets to carry out pre-training and fine-tuning of the models. Since the pre-training is unsupervised, it does not require a labeled dataset. However, it required two separate datasets annotated with four classes (Positive, Negative, Neutral, and Conflict) and three classes (Positive, Negative, and Neutral) to fine-tune the models.

4.1 Dataset for pre-training

We used the Sinhala corpus extracted from OSCAR dataset to pre-train the models. OSCAR dataset is a multilingual corpus obtained by language classification and filtering of the Common Crawl corpus using the Ungoliant architecture. Common Crawl corpus is a huge corpus that contains petabytes of raw web page data, metadata extracts, and text extracts gathered over 12 years of web crawling [31]. The OSCAR dataset has raw text from 162 languages, including the Sinhala language. This dataset contains 108,593 documents in the Sinhala language and 113,179,741 Sinhala words. The total size of the dataset is around 2.0 GB [31].

4.2 Dataset for fine-tuning

The dataset published by Senevirathne et al. [30] in 2020 contains 15059 News comments annotated with four classes: Positive, Negative, Neutral, and Conflict. This dataset contains 9059 News comments from the dataset published by

Ranathunga and Liyanage [6] and another 6000 News comments extracted from GossipLanka News websites. This annotation has been done by three annotators following the guidelines mentioned below [30],

- A comment is annotated as positive or negative if it expresses a purely positive or negative opinion.
- A comment is annotated as a conflict if it gives both positive and negative opinions.
- A comment is annotated as neutral if it does not give any positive or negative opinion.

In this study, we expanded this dataset by following the steps below.

Data collection: We collected 803,623 news comments from the GossipLanka News website and filtered the comments posted only using Sinhala Unicode characters (Range: 0D80 - 0DFF). There were 417,332 comments posted using Sinhala Unicode.

Data annotation: Two annotators who are native Sinhala speakers carried out the annotating task following the guidelines mentioned previously. We used Cohen's Kappa measure to evaluate the inter-annotator agreement, which yielded a value of 0.794. Both annotators collectively annotated 5,037 Sinhala News comments with four classes (Positive, Negative, Neutral, and Conflict). These annotated comments were added to the existing Sinhala News comments dataset. We carefully removed duplicate entries from the resulting dataset to ensure data integrity. The final dataset comprised 19,875 unique comments.

Classes	Dataset 1 (Four Classes)	Dataset 2 (Three Classes)
Positive	3,587	4,414
Negative	10,228	11,639
Neutral	3,822	3,822
Conflict	2,238	0
Total	19,875	19,875

Table 1: Distribution of comments per class

The newly generated dataset was annotated using Positive, Negative, and Neutral to create a sentiment dataset with three classes. Comments initially labeled as Conflict were annotated as Positive or Negative based on their predominant sentiment. Table 1 shows the distribution of comments per class in the two datasets.

4.3 Model pre-training

It was not necessary to pre-train the XLM-R-base and XLM-R-large models for the Sinhala language since it is already pre-trained on a multilingual corpus that includes Sinhala. However, we had to pre-train the other three models for the Sinhala language, and those three models were pre-trained using the Sinhala dataset extracted from the OSCAR dataset.

Since models cannot process raw data directly, they need to be converted to a representation that the models can process. Therefore, it was necessary to train tokenizers for these models from scratch. BERT and DistilBERT tokenizers use the WordPiece method [4, 13], while the RoBERTa tokenizer uses the Byte-Pair Encoding method [15]. Tokenizers for the three models were trained with a vocabulary size of 52,000 and a minimum frequency of 2 using the Sinhala dataset extracted from the OSCAR dataset. The vocabulary size defines the number of all tokens and alphabets included in the final vocabulary, and the minimum frequency defines the minimum frequency a pair should have to be merged. Special tokens included in BERT, DistilBERT, and RoBERTa tokenizers are listed in Table 2.

BERT	DistilBERT	RoBERTa
<s>	<s>	[PAD]
<pad>	<pad>	[UNK]
</s>	</s>	[CLS]
<unk>	<unk>	[SEP]
<mask>	<mask>	[SEP]

Table 2: Special tokens included in tokenizers

After training the tokenizers, the three models were built using the parameters listed in Table 3. The max position embedding column shows the maximum sequence length that this model can use, and the dimensionality of the encoder layers and the pooler layers is denoted by the hidden size column. The last two columns show the number of attention heads for each attention layer and the number of hidden layers. After building the models, they were trained for masked language modeling task using the same dataset used to train the tokenizers. We used AdamW as the optimizer with a learning rate 5e-5 and a batch size 16. Models were only trained for one epoch as pre-training is computationally expensive.

Parameters	BERT	DistilBERT	RoBERTa
Vocabulary Size	52,000	52,000	52,000
Max Position Embedding	512	512	512
Hidden Size	768	768	768
Attention Heads	12	12	12
Hidden Layers	12	12	12

Table 3: Parameters used for building the models

Models	Four Classes				Three Classes			
	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall
BERT	0.4674	0.4898	0.4719	0.5173	0.5829	0.6254	0.5743	0.6114
DistilBERT	0.4704	0.4943	0.4730	0.5129	0.5699	0.6058	0.5641	0.6015
RoBERTa	0.4270	0.4453	0.4292	0.4729	0.5146	0.5585	0.5127	0.5429
XLM-R$_{base}$	0.5836	0.6171	0.5825	0.6183	0.6808	0.7175	0.6673	0.7052
XLM-R$_{large}$	0.6344	0.6719	0.6256	0.6622	0.7135	0.7301	0.6827	0.6906

Table 5: Results for sentiment analysis using four classes and three classes

Parameters	BERT	DistilBERT	RoBERTa	XLM-R$_{base}$	XLM-R$_{large}$
Batch Size	16	16	16	16	16
Dropout Rate	0.1	0.1	0.1	0.1	0.1
Learning Rate	2e-5	2e-5	1e−5	5e−6	5e−6
Weight Decay	0.01	0.01	0.01	0.01	0.01
Epochs	5	5	10	5	5
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW

Table 4: Parameters used for fine tuning the models

4.4 Model fine-tuning

The pre-trained models should be fine-tuned to carry out sentiment analysis. Even though XLM-R-base and XLM-R-large are already pre-trained for the Sinhala language, it needs to be fine-tuned for sentiment analysis in the Sinhala language. Therefore, all five pre-trained models were fine-tuned for sentiment analysis. Each pre-trained model was fine-tuned twice using Dataset 1 and Dataset 2 separately. According to the original paper of BERT [4], the recommended number of epochs for fine-tuning a model is 2, 3 and 4. Therefore BERT and DistilBERT models were fine-tuned for five epochs and at the end of each epoch, the trained model was saved as a checkpoint. The best performing model was picked from the saved checkpoints by considering the loss at each epoch. The parameters used for fine-tuning the models are listed in Table 4.

5 Results and Discussion

We evaluated the performance of the fine-tuned models using accuracy, macro-F1 score, macro-precision, and macro-recall. The results obtained by Dhananjaya et al. [x] for the sentiment task serve as the baseline for our study. Table 5 presents the results obtained for sentiment analysis for three and four classes. In this study, we conducted all model training and evaluation using the Transformers library provided by HuggingFace on the Google Colab Pro environment.

For sentiment analysis using four classes, we observe that XLM-R-large achieved the highest macro-F1 score of 0.6344, followed closely by XLM-R-base with a macro-F1 score of 0.5836. Similarly, XLM-R-large continues to display superior performance for sentiment analysis using three classes, achieving a macro F1-score of 0.6808 and an accuracy of 0.7175. Dhananjaya et al. [x] obtained a macro-F1 score of 0.6345 for sentiment analysis using four classes, which serve as the baseline model. This indicates that our model outperformed the baseline model slightly. One potential reason for the improved performance of XLM-R-large is the utilization of a larger training dataset, allowing the model to learn from a more diverse set of examples and generalize better.

Our study observed that BERT and DistilBERT achieved competitive macro-F1 scores and accuracy for both sentiment analysis tasks with four and three classes. However, the macro-F1 scores of BERT, DistilBERT, and RoBERTa were relatively lower than XLM-R models. The outcome of these monolingual models achieving lower results than XLM-R models was unexpected. Monolingual models are typically trained specifically for a single language, and they would have a better understanding of linguistic patterns, leading to better performance in sentiment analysis tasks. However, the observed results highlighted that XLM-R models performed better in sentiment analysis for Sinhala despite being pre-trained on a multilingual corpus. The reason for this unexpected outcome is the difference in the pre-training process. BERT, DistilBERT, and RoBERTa models were pre-trained for only one epoch, while XLM-R models were pre-trained for a higher number of epochs. This longer pre-training process allowed XLM-R models to gain a deeper understanding of linguistic patterns and representations, making them more effective in sentiment analysis for Sinhala. However, it is important to note that these monolingual models still demonstrate promising capabilities in capturing sentiment patterns in Sinhala text. The performance of these monolingual models can be further improved by pre-training the models on a larger Sinhala corpus for a higher number of epochs.

Figure 1 displays the confusion matrix for sentiment analysis conducted using the XLM-R-large model with four classes. Based on the confusion matrix, we can deduce that the XLM-R-large model performs better in predicting the majority classes (Negative, Neutral, and Positive) than the Conflict class. There is a noticeable tendency for the model to misclassify instances labeled as Conflict as Negative at a relatively higher frequency. This misclassification pattern may be influenced by the class imbalance in the dataset, where the Negative class is the majority class with over 10,000 instances. The model might have learned to favor the majority class, leading to more frequent misclassifications for the Conflict class. The class imbalance poses a challenge for the model to accurately distinguish between the classes, particularly affecting its ability to predict the minority class accurately.

6 Conclusion and Future Work

This study evaluates the performance of various transformer models fine-tuned for sentiment analysis in the Sinhala language. This study marks the first experimentation of BERT and DistilBERT for sentiment analysis in Sinhala. The findings demonstrate that transformer models exhibit remarkable performance, even when fine-tuned using a small dataset. This outcome highlights the significant potential of transformer models in addressing challenges for languages with limited available resources. We also showed that the extensive pre-training process of the XLM-R models played a pivotal role in their superior performance compared to other models pre-trained for a single epoch.

In this study, we have made several contributions to the research community. We have made publicly available the pre-trained models of BERT, DistilBERT, and RoBERTa, along with the fine-tuned models of BERT, DistilBERT,

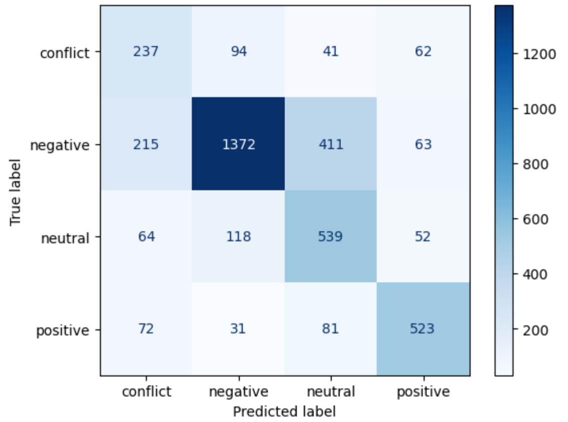

Figure 1: Confusion matrix of XLM-R-large for sentiment analysis with four classes.

RoBERTa, XLM-R-base, and XLM-R-large. Additionally, three datasets have been released, which include a sentiment dataset comprising 19,875 news comments annotated with four classes, another dataset with 19,875 news comments annotated for three classes, and a large Sinhala news comments dataset containing 417,332 unannotated comments. These resources aim to foster further advancements and enable researchers to explore sentiment analysis in the Sinhala language more effectively.

These research outcomes contribute valuable insights to the field of sentiment analysis of low-resource languages and provide a foundation for future studies and applications. The utilization of transformer models, especially XLM-R-large, showcased promising results, indicating the potential for further advancements in sentiment analysis tasks for the Sinhala language.

In the future, we plan to explore the performance of other transformer models, such as ALBERT and GPT-2, for sentiment analysis in the Sinhala language.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the*

Association for Computing Machinery, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. *Yara parser: A fast and accurate depen-dency parser. Computing Research Repository,* arXiv:1503.06733. Version 2

A Appendices

Appendices are added after the References section by restarting the header numbering using style “A, B, C”.

B Supplementary Material

Supplementary material also be included with the Appendices.