A Dual Contrastive Learning Framework for Enhanced Hate Speech Detection in Low-Resource Languages

Krishan Chavinda
Department of Computer Science
& Engineering
University of Moratuwa
Sri Lanka
krishan.19@cse.mrt.ac.lk

Uthayasanker Thayasivam
Department of Computer Science
& Engineering
University of Moratuwa
Sri Lanka
rtuthaya@cse.mrt.ac.lk

Abstract

Hate speech on social media platforms is a critical issue, especially in low-resource languages such as Sinhala and Tamil, where the lack of annotated datasets and linguistic tools hampers the development of effective detection systems. This research introduces a novel framework for detecting hate speech in low resource languages by leveraging Multilingual Large Language Models (MLLMs) integrated with a Dual Contrastive Learning (DCL) strategy. Our approach enhances detection by capturing the nuances of hate speech in low-resource settings, applying both self-supervised and supervised contrastive learning techniques. We evaluated our framework using datasets from Facebook and Twitter, demonstrating its superior performance compared to traditional deep learning models such as CNN, LSTM, and BiGRU. The results highlight the efficacy of DCL models, particularly when fine-tuned on domain-specific data, with the best performance achieved using the TwHIN-BERT model. This study underscores the potential of advanced machine learning techniques in improving hate speech detection for under-resourced languages, paving the way for further research in this domain.

1 Introduction

Hate speech has become a significant problem in the digital age, particularly on social media platforms where communication is fast, widespread, and often anonymous. The rise of online platforms has not only enhanced global connectivity but has also provided a fertile ground for the dissemination of harmful content, including hate speech.We adopt the definition by Lu et al. (2023): “Hate speech is subjective and derogatory speech towards protected characteristics expressed directly or indirectly to such groups in textual form”. This form of expression, which targets individuals or groups based on attributes such as race, religion, ethnicity, gender, or sexual orientation, has severe societal impacts, including the escalation of violence, promotion of discrimination, and deepening of social divides. The urgency to address hate speech is further underscored by its ability to rapidly spread and amplify through social networks, making it a formidable challenge for regulatory bodies and social media companies alike.

In multicultural societies, hate speech has been a particular concern due to its potential to incite ethnic and religious violence. However, while considerable research has been devoted to hate speech detection in languages like English, less attention has been given to low-resource languages such as Sinhala and Tamil. The lack of resources, such as annotated datasets and linguistic tools, has posed significant challenges to developing robust hate speech detection systems in low-resource languages.

Despite the growing interest in using advanced technologies to combat hate speech, the application of Large Language Models (LLMs) in this domain remains underexplored, particularly in the context of Sinhala. LLMs, with their ability to understand and generate human-like text, offer a promising avenue for improving hate speech detection. However, existing studies (Munasinghe and Thayasivam, 2022) (Samarasinghe et al., 2020) (Hettiarachchi et al., 2020) (Sandaruwan et al., 2019) in Sinhala have predominantly relied on traditional machine learning models, which, while effective, may not capture the nuances of the language or the context of the speech as effectively as LLMs. Therefore, this research aims to bridge this gap by exploring the potential of LLMs for detecting hate speech in Sinhala as well as providing a different approach for hate speech detection in low resource languages, focusing specifically on content generated on social media platforms.

This study aims to contribute to the field by developing and evaluating a hate speech detection model for Sinhala and Tamil, two low-resource languages, utilizing the capabilities of large language

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 115–123
January 19, 2025.

models (LLMs). The findings of this research could provide valuable insights into the effectiveness of LLMs in low-resource languages and offer a foundation for future work in this critical area of study.

2 Related Work

We have conducted an extensive literature review, primarily focused on hate speech detection in Sinhala, a low-resource language, as our main focus lies within this linguistic domain. The research into hate speech detection in the Sinhala language, particularly in the context of social media, has garnered significant attention due to the growing prevalence of online abusive content. Various studies have employed different machine learning techniques and natural language processing (NLP) methods to address this issue, highlighting the unique challenges presented by the Sinhala language and its Romanized form.

In the study by Munasinghe and Thayasivam (2022), a deep learning ensemble method was introduced to detect hate speech in Sinhala tweets. Their approach contrasts with previous models that primarily relied on traditional machine learning methods like Naive Bayes, Support Vector Machines (SVM), and Random Forest classifiers, which often struggled to generalize due to limited dataset sizes and suboptimal results. By creating a new dataset using Twitter API and applying techniques such as stop word removal, stemming, and tokenization, they were able to develop a deep learning model based on convolutional neural networks (CNN), Long Short-Term Memory (LSTM), and Bi-GRU models. This ensemble method yielded superior performance, achieving over 90% accuracy, precision, recall, and F-scores. The ensemble's ability to outperform individual models demonstrates the potential of deep learning for hate speech detection in low-resource languages like Sinhala.

The research by Samarasinghe et al. (2020) focused on the detection of hate speech in Sinhala Unicode text, utilizing a CNN model with FastText word embeddings. This study introduced a two-stage classification process where the first step identified hate speech, and the second classified it according to the severity of the hate. While the study achieved high accuracy in the hate speech classification (83%), it highlighted the challenges in identifying varying levels of hate speech, particularly due to the imbalanced dataset. The difficulty in accessing larger, more diverse datasets further limited the model's ability to generalize, underlining a significant barrier in hate speech detection for Sinhala.

Hettiarachchi et al. (2020) extended the scope of hate speech detection by focusing on Romanized Sinhala, a unique form of Sinhala written using the English script. Their research applied a variety of machine learning algorithms, including Logistic Regression, Naive Bayes, SVM, and Random Forest, to a dataset of Facebook comments written in Romanized Sinhala. Despite the language's complexities, including inconsistencies in spelling and grammar, the study found that the Naive Bayes classifier performed best with bigram features, achieving an accuracy of 71%. This research emphasized the potential of applying machine learning methods to non-standard linguistic forms, particularly for low-resource languages like Sinhala.

Further research conducted by Dias et al. (2018) explored racist comments in Sinhala social media using text analytics models. They experimented with a Support Vector Machine (SVM) classifier using a set of Facebook comments labeled as either racist or non-racist. Their results, with a 70.8% accuracy, highlighted the challenges of detecting racist comments specifically, which share many linguistic traits with general offensive comments. The imbalanced dataset and the difficulty of separating intent from context were key hurdles. The study recommended that future research use more sophisticated NLP techniques to improve detection rates.

In a similar vein, Fernando and Deng (2023) introduced a novel approach that enhanced hate speech detection in Sinhala by applying feature selection techniques. Their study proposed a global feature selection process to tackle high-dimensional input data challenges, using classifiers such as SVM, Multinomial Naive Bayes (MNB), and Random Forest. The research demonstrated that advanced feature selection could significantly improve detection performance in the sparse and noisy datasets typical of Sinhala social media, particularly when combined with character and word n-grams. This approach also revealed improvements in model generalization across training and testing datasets.

Moreover, Sandaruwan et al. (2019) explored the lexicon-based and machine learning approaches for hate speech detection in Sinhala social media, where they used a corpus of 3,000 comments. Their study revealed that the Multinomial Naive Bayes classifier, when combined with character trigrams, achieved the highest accuracy of 92.33%. This lexicon-based approach also showed promise in identifying hate, offensive, and neutral speech categories, though the study underlined the need for better feature engineering and larger datasets to improve the model's scalability.

Across these studies, common themes emerge, including the importance of dataset size, feature engineering, and model selection. Traditional machine learning models, while effective in certain cases, struggle with generalization when faced with small or imbalanced datasets, as evidenced by the drop in performance in various studies. Deep learning models, particularly those that leverage ensemble techniques, have demonstrated more robust performance, although they require significantly more data and computational resources. Moreover, the detection of hate speech in Romanized Sinhala adds another layer of complexity, necessitating the exploration of feature extraction methods that can handle linguistic variations. In conclusion, the advancement of hate speech detection in the Sinhala language relies heavily on the availability of large, annotated datasets and the continued development of sophisticated NLP models.

3 Methodology

In this research, we introduce a novel framework designed specifically for hate speech detection in low-resource languages, by leveraging multilingual Large Language Models (MLLMs). This framework adapts and extends the Dual Contrastive Learning (DCL) strategy proposed by Lu et al. (2023), integrating enhancements suitable for handling the nuances of low resource language hate speech on social media platforms.

3.1 Dual Contrastive Learning Framework for Low Resource Languages

The framework depicted in figure 1 represents a novel Dual Contrastive Learning (DCL) approach specifically tailored for hate speech detection in low resource languages, leveraging multilingual Large Language Models (MLLMs) that support low-resource languages. The overall framework involves the following steps:

1. Embedding Generation with Multilingual LLM: Input sentences, including both hate and non-hate speech, are processed using a pre-trained multilingual LLM that supports low-resource languages. This model generates contextual embeddings that capture the semantic meaning of the input text.

2. Data Augmentation through Dropout: To enhance the training data, dropout-based data augmentation is applied to the embeddings generated by the LLM. This process creates multiple augmented views of each input, which are used in subsequent contrastive learning stages.

3. Dual Contrastive Learning Mechanisms: The proposed framework employs two stages of contrastive learning:
    - **Self-Supervised Contrastive Learning:** This stage focuses on learning invariant representations by creating positive pairs from augmented views of the same hate speech sample. Strong data augmentation techniques are used to generate these pairs, aiming to maximize the separation between these positive pairs and negative pairs (non-hate speech) in the embedding space.
    - **Supervised Contrastive Learning:** This stage utilizes label information to refine the representation space by pulling samples from the same class closer together while pushing apart those from different classes. This clustering effect improves the model's ability to distinguish between hate and non-hate speech effectively.

The integration of these stages allows the framework to capture both intrinsic patterns within hate speech and the discriminative features between hate and non-hate content, thereby enhancing the detection capabilities for low resource language hate speech on social media platforms. For the above learning, there will be losses to identify the performance of the Model.

3.2 Self-Supervised Contrastive Learning

Considering the complexity and ambiguity of hate speech expressions, we use self-supervised contrastive learning for data augmentation and deeper semantic feature extraction. By constructing positive and negative

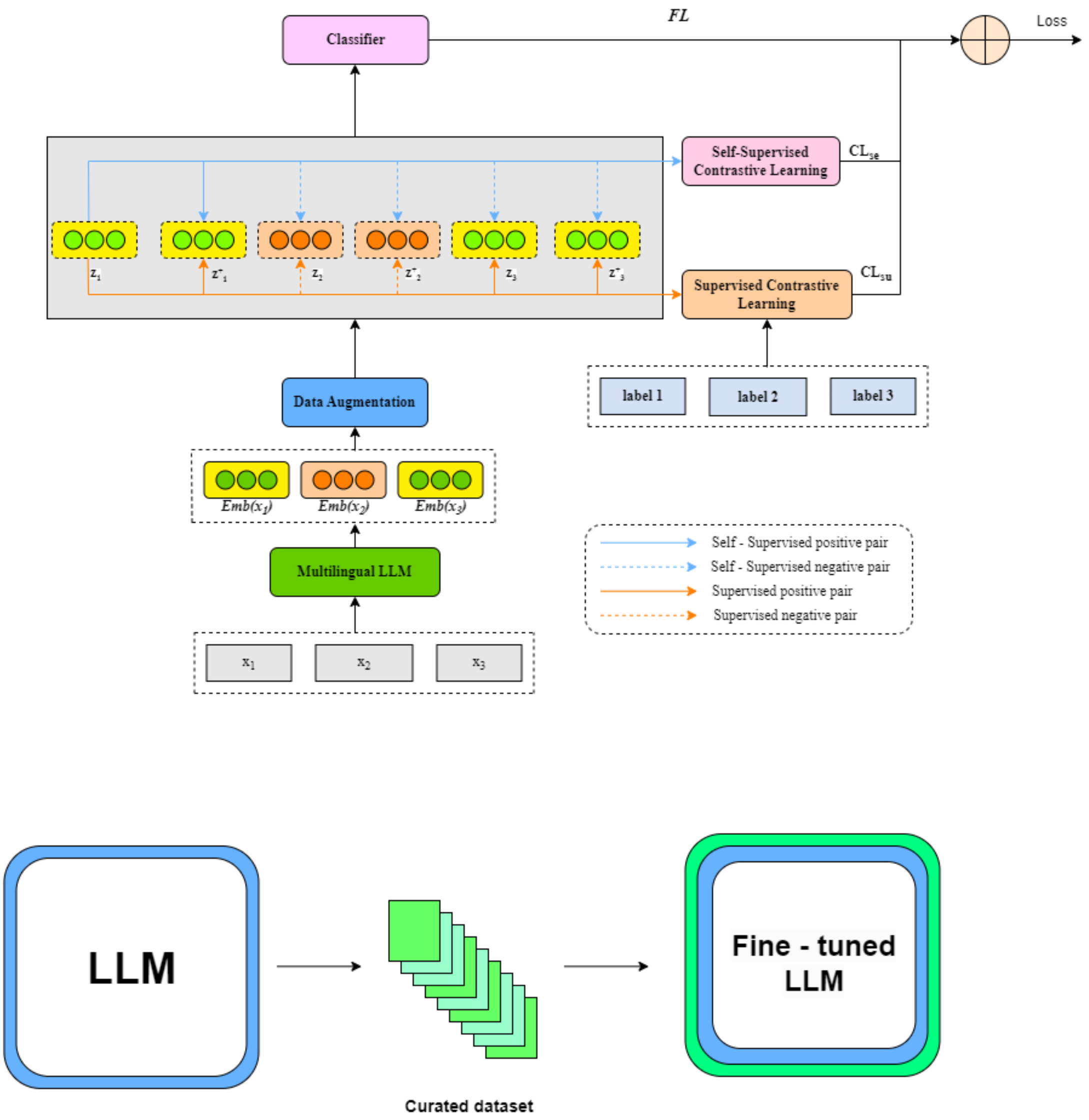

Figure 1: Dual Contrastive Learning Framework for Low Resource Languages

.

samples, this approach captures more comprehensive span-level features, going beyond token-level semantics, to better distinguish subtle differences (Gao et al., 2021).

$$\mathrm{CL}_{se} = -\sum_{j=1}^{2N} log \frac{e^{sim(z_j, z_j^+)/\Gamma_{se}}}{\sum_{k=1}^{2N} 1_{[j \neq k]} \cdot e^{sim(z_j, z_k)/\Gamma_{se}}}, \quad (1)$$

CL_{se} represent the Self-Supervised Contrastive Learning loss and for a given input sentence x_i, standard dropout is applied twice to create the sentence embedding $Emb(x_i)$ that retains maximum semantic information (Srivastava et al., 2014). This embedding is then used to generate two augmented samples, z_j and z_j^+, through independently sampled dropout masks on fully-connected layers. The pair (z_j, z_j^+) serves as positive samples, while other samples in the batch are treated as negatives. The parameters include N for batch size before data augmentation and τ_{se} as a non-negative temperature hyperparameter. The function $sim(\cdot)$ calculates the similarity scoring between z_j and z_j^+ using cosine similarity to guide the contrastive objective, encouraging similar embeddings for augmented variants of the same input and contrasting them against others.

3.3 Supervised Contrastive Learning

To improve hate speech detection, we first apply self-supervised contrastive learning to highlight important span-level semantics within the data. Next, we incorporate label information through supervised contrastive learning. This approach ensures that examples sharing the same label (positive samples) are drawn closer together in the embedding space, while those with different labels (negative samples) are pushed apart. By doing so, the model benefits from both the self-supervised augmentation and the explicit label guidance.

Given a batch of N samples, the supervised contrastive loss CL_{su} is defined as:

$$\mathrm{CL}_{su} = -\textstyle\sum_{i=1}^{N} \frac{1}{N_{y_i}-1} \sum_{j=1}^{N} 1_{[i \neq j]} \cdot 1_{[y_i \neq y_j]}$$
$$\cdot \log \frac{e^{\mathrm{sim}(z_j, z_j)/\Gamma_{su}}}{\sum_{k=1}^{N} 1_{[i \neq k]} e^{\mathrm{sim}(z_j, z_k)/\Gamma_{su}}}$$

(2)

Here, (z_i, z_j) is a pair of positive samples (with the same label),and (z_i, z_k) represents a comparison to a randomly chosen sample.The labels of z_i and z_j are denoted by y_i and y_j, respectively, with N_{y_i} representing the count of samples sharing the same label as z_i. The non-negative temperature coefficient Γ_{su} modulates the supervised contrastive loss.

3.4 Dual Contrastive and Focal losses Integration

To jointly integrate both self-supervised and supervised signals, we define our overall loss function for contrastive learning as follows:

$$CL = CL_{se} + CL_{su} \tag{3}$$

Dual contrastive learning objectives (losses) are then integrated with the focal loss(Ross and Dollár, 2017) function which addresses data imbalance issues in hate speech detection to obtain the total loss function, which will be optimized to obtain the fine-tuned DCL model.

The Focal loss is defined as follows:

$$FL = -\sum_{i=1}^{N} \alpha_i (1 - \hat{p_i})^{\gamma} log(\hat{p_i}) \tag{4}$$

The parameter γ, a non-negative tuning factor, distinguishes between easy and challenging samples in the context of model's learning. A lower γ encourages the model to prioritize misclassified instances, diminishing the impact of well-classified samples. Additionally, α, ranging from 0 to 1, serves as a weighting factor, ensuring a balance in the significance attributed to positive and negative samples, which is defined as,

$$\alpha_i = \begin{cases} \alpha, & \text{if } y_i = 1 \\ 1 - \alpha, & \text{otherwise} \end{cases} \tag{5}$$

$\hat{p_i}$ in (4) reflects the relationship between the estimated probability and the target class.

$$\hat{p_i} = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases} \tag{6}$$

$p_i \in$ [0, 1] is the estimated probability for the class with the label y_i = 1 in each sentence embedding z_i.

This adaptive approach enhances the model's ability to focus on challenging instances and effectively balances the influence of different sample types in the learning process.

The Total Loss function is defined as follows:

$$Loss = FL + \lambda \cdot CL \tag{7}$$

A weighting coefficient λ is used to balance the impact of these two losses, where $\lambda \in$ [0, 1].

4 Experiments

In this section, we evaluate the performance of our DCL Framework for low-resource languages using two multilingual LLMs: xlm-roberta-base (Conneau et al., 2019) and twhin-bert-base (Zhang et al., 2022), both of which support Sinhala and Tamil. We introduce three publicly available datasets, describe our experimental setups, and present evaluation results that compare our model with other baseline deep learning models. We then analyze these results in detail.

For hyperparameter tuning, we employed Optuna (Akiba et al., 2019), and for improved experiment tracking and monitoring, we incorporated Neptune.ai. Our experiments primarily focus on the Sinhala language, providing a comparative analysis of model performance. We also conducted experiments for the Tamil language, but without a comparative perspective.

4.1 Datasets

For our research, we used the following three publicly available datasets. [1]

4.1.1 Facebook Sinhala Hate Speech Dataset

This dataset contains a total of 6,345 samples, sourced from Facebook. It features a near-balanced distribution with 3,455 instances of hate speech (54.45%) and 2,890 instances of non-hate speech (45.55%). The balanced nature of this dataset, combined with its real-world context from Facebook, makes it highly applicable for developing and assessing hate speech detection models tailored for social media platforms. Its comprehensive representation of both hate and non-hate speech ensures that models trained on this data can effectively generalize to similar scenarios on Facebook.

4.1.2 Twitter Sinhala Hate Speech Dataset

Comprising 4,502 samples collected from Twitter, this dataset includes 1,108 instances of hate speech (24.62%) and 3,394 instances of non-hate speech (75.38%). The dataset's higher proportion of non-hate speech mirrors the typical distribution on Twitter, providing valuable insights for detecting hate speech in real-world social media environments. Its focus on the Twitter platform allows for effective training and evaluation of hate speech detection models, particularly in handling imbalanced data and adapting to the nuances of Twitter's social media interactions.

4.1.3 Tamil Hate Speech Dataset

The Tamil hate speech dataset consists of a total of 5,503 labeled instances, with 3,573 classified as non-hate speech and 1,930 as hate speech. This distribution indicates that approximately 64.9% of the dataset is non-hate speech, while 35.1% is hate speech. The dataset is slightly imbalanced, with a higher proportion of non-hate speech compared to hate speech, though the imbalance is not extreme.

4.2 Experimental Settings

In this section, we describe the experimental setup for our framework. We conducted experiments using two multilingual large language models (LLMs) integrated with our framework and evaluated their performance on the datasets. These experiments demonstrate that our approach outperforms traditional state-of-the-art deep learning methods. The datasets were divided into training and test sets, and we employed 5-fold cross-validation for each dataset to assess model performance on the test set.

For training, we used a dropout rate of 0.5 and the AdamW optimizer (Kingma and Ba, 2014). Hyperparameter tuning was performed to optimize the batch size, learning rate, number of epochs, and additional parameters such as λ, τ_{se}, τ_{su}, as well as the focal loss parameters α and γ.

Model performance was primarily assessed using the weighted F1 score with cross-validation. We selected the models and hyperparameters that achieved the best validation results and further evaluated these on the test set using metrics such as accuracy, weighted F1 score, precision, and recall.

All models were trained on an NVIDIA T4 GPU to ensure efficient computation.

5 Results & Analysis

The performance metrics of several deep learning models, including CNN, LSTM, BiGRU, an ensemble of these models, and DCL models($DCL_{XLM-RoBERTa}$ and $DCL_{TwHIN-BERT}$), were evaluated on two sinhala datasets: the Twitter Sinhala Hate Speech Dataset and the Facebook Sinhala Hate Speech Dataset. The results are presented in Table 1 and Table 2. In addition, the results presented in Table 3 show the performance of the DCL models on the Tamil Hate Speech Dataset.

Based on the results, our $DCL_{TwHIN-BERT}$ model outperformed on both Sinhala datasets (Table 1 and Table 2). This highlights the importance of employing advanced machine learning techniques to address challenges in hate speech detection within under-resourced language contexts.

The $DCL_{TwHIN-BERT}$ model achieved the highest performance on the Twitter Sinhala Hate Speech Dataset, with 94.00% accuracy and balanced F1, recall, and precision scores, highlighting its robustness. The $Ensemble_{(CNN,LSTM,BiGRU)}$ model slightly surpassed in accuracy (94.10%) but underperformed in F1 score, recall and precision, suggesting less balanced generalization. On the Facebook Sinhala Hate Speech Dataset, the $DCL_{TwHIN-BERT}$ model again excelled, achieving 87.29% accuracy, outperforming the ensemble model (85.70%) and other models.

[1] Here are the publicly available Datasets: (1) Facebook Sinhala Hate Speech (2) Twitter Sinhala Hate Speech (3) Tamil Hate Speech

Table 1: Performance Metrics for Deep Learning Models on the Twitter Sinhala Hate Speech Dataset

Model	**Accuracy**	**F1 Score**	**Recall**	**Precision**
CNN[2]	90.10%	90.10%	90.10%	90.10%
LSTM[2]	91.90 %	91.90%	91.90%	91.90%
BiGRU[2]	92.80%	92.80%	92.80%	92.80%
$Ensemble_{(CNN,LSTM,BiGRU)}$[2]	**94.10%**	91.90%	93.00%	90.10%
$DCL_{XLM-RoBERTa}$	91.50%	91.60%	91.50%	91.90%
$DCL_{TwHIN-BERT}$	94.00%	**94.00%**	**94.00%**	**94.00%**

[2] denotes results obtained from the literature

Table 2: Performance Metrics for Deep Learning Models on the Facebook Sinhala Hate Speech Dataset

Model	**Accuracy**	**F1 Score**	**Recall**	**Precision**
CNN[3]	84.80 %	84.80%	84.80%	84.80%
LSTM[3]	83.10%	83.10%	83.10%	83.10%
BiGRU[3]	85.10%	85.10%	85.10%	85.10%
$Ensemble_{(CNN,LSTM,BiGRU)}$[3]	85.70%	85.70%	85.70%	85.70%
$DCL_{XLM-RoBERTa}$	85.82%	85.85%	85.82%	86.08%
$DCL_{TwHIN-BERT}$	**87.29%**	**87.19%**	**87.29%**	**87.61%**

[3] denotes results obtained from the literature

5.1 Key Observations

- $DCL_{TwHIN-BERT}$ model was the top performer across all datasets, highlighting the power of DCL-based approach in detecting hate speech on low resource languages.

- $Ensemble_{(CNN,LSTM,BiGRU)}$ model showed competitive results, particularly on the Twitter dataset, but were outperformed by the DCL model, especially in terms of F1 score, recall, and precision.

- The traditional models (CNN, LSTM, BiGRU) demonstrated reasonable performance but did not reach the level of the DCL models, emphasizing the importance of pre-trained language models fine-tuned for specific tasks such as hate speech detection.

5.2 Performance Comparison between XLM-RoBERTa and TwHIN-BERT based DCL Models

The XLM-RoBERTa model is pre-trained on the CommonCrawl Corpus (CC-100), which primarily comprises data collected from open web pages (Conneau et al., 2019). In particular, this corpus excludes social media data. Consequently, the model lacks exposure to social media-specific patterns, trends, and expressions, which are often informal, context-specific, and culturally nuanced. Furthermore, low-resource languages such as Sinhala and Tamil have limited representation in digital content (Joshi et al., 2020), making it challenging for the model to capture the unique linguistic characteristics of these languages as they appear in social media contexts.

We hypothesize that this limitation in XLM-RoBERTa's pre-training significantly restricts the potential performance gains achievable through its integration with the DCL framework. This stands in contrast to the DCL framework employing TwHIN-BERT, a model pre-trained on 7 billion tweets across 100 languages. TwHIN-BERT leverages textual data alongside social engagement signals through the Twitter Heterogeneous Information Network (TwHIN) (Zhang et al., 2022). This socially enriched pretraining enables TwHIN-BERT to better understand the informal and context-specific linguistic expressions prevalent on social media platforms.

Given these differences, our results demonstrate that the $DCL_{TwHIN-BERT}$ model excels in hate speech detection, which requires social media-specific linguistic understanding. The inclusion of social media data during TwHIN-BERT's pretraining enables it to capture cultural nuances and informal language variations more effectively than XLM-RoBERTa. This advantage is reflected in the observed superior performance of TwHIN-BERT-based implementations compared to their XLM-RoBERTa counterparts in hate speech detection for

Table 3: Performance Metrics for Deep Learning Models on a Tamil Hate Speech Dataset

Model	Accuracy	F1 Score	Recall	Precision
$DCL_{XLM-RoBERTa}$	65.86%	56.32%	65.86%	63.86%
$DCL_{TwHIN-BERT}$	74.58%	74.38%	74.58%	74.26%

low-resource languages on social media.

5.3 Limitations

Our work primarily focuses on hate speech in the Sinhala language, with limited exploration of Tamil among low-resource languages. Experiments were conducted exclusively with XLM-RoBERTa and TwHIN-BERT models, leaving scope for future exploration of other multilingual large language models (MLLMs).

Conclusion

This study highlights the potential of advanced machine learning techniques, particularly use of dual contrastive learning with pre-trained multilingual LLMs like XLM-RoBERTa and TwHIN-BERT, for hate speech detection in low-resource languages such as Sinhala and Tamil.Our DCL framework-based model outperformed existing state-of-the-art traditional deep learning models, with the TwHIN-BERT-based DCL model consistently achieving superior performance across both Sinhala datasets.In addition, our findings reveal the critical importance of domain-specific pretraining on social media data, as demonstrated by TwHIN-BERT, in addressing the challenges of informal and context-dependent expressions prevalent on social media platforms, particularly for hate speech detection in low-resource languages. These results lay a strong foundation for future research in hate speech detection for low-resource languages using Multilingual Large Language Models.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Dulan S Dias, Madhushi D Welikala, and Naomal GJ Dias. 2018. Identifying racist social media comments in sinhala language using text analytics models with machine learning. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 1–6. IEEE.

Eranga N Fernando and Jeremiah D Deng. 2023. Enhancing hate speech detection in sinhala language on social media using machine learning.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Nimali Hettiarachchi, Ruvan Weerasinghe, and Randil Pushpanda. 2020. Detecting hate speech in social media articles in romanized sinhala. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 250–255. IEEE.

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. *arXiv preprint arXiv:2004.09095*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. 2023. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sidath Munasinghe and Uthayasanker Thayasivam. 2022. A deep learning ensemble hate speech detection approach for sinhala tweets. In *2022 Moratuwa Engineering Research Conference (MERCon)*, pages 1–6. IEEE.

T-YLPG Ross and GKHP Dollár. 2017. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988.

SWAMD Samarasinghe, RGN Meegama, and M Punchimudiyanse. 2020. Machine learning approach for the detection of hate speech in sinhala unicode text. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 65–70. IEEE.

HMST Sandaruwan, SAS Lorensuhewa, and MAL 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. IEEE.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Xinyang Zhang, Yury Malkov, Omar U. Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.