

Enhancing Hate Speech Detection in Sinhala and Gujarati: Leveraging BERT Models and Linguistic Constraints

G Gana Sai^{1,*†}, Aswath Venkatesh^{1,†}, Kishore N^{1,†}, Odirva M^{1,†}, Balaji V A^{1,†} and Prabavathy Balasundaram^{2,†}

¹UG Student, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India.

²Faculty, Department of Computer Science, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract

This research paper, presented by the SSN_CSE_ML_TEAM, introduces a unified approach to hate speech and offensive language identification in two low-resource Indo-Aryan languages, Sinhala and Gujarati, as part of the HASOC 2023 shared tasks. Leveraging various BERT models, we address the challenge of classifying tweets into Hate and Offensive (HOF) and Non-Hate and Offensive (NOT) categories by fine-tuning the BERT models. Our approach seeks to advance the state-of-the-art in detecting hate speech while considering the unique linguistic characteristics and resource constraints of these languages.

Keywords

Hate Speech Detection, Offensive Language Identification, BERT Models, Text Classification, Multilingual NLP

1. Introduction

Online communication platforms have become integral to modern society, enabling diverse linguistic communities to interact and express their thoughts and opinions. However, these platforms are not immune to the proliferation of hate speech and offensive language, which can have severe social and psychological consequences. Addressing this issue is of utmost importance, and it becomes particularly challenging in low-resource languages where language-specific resources and models are limited.

This research paper tackles the problem of hate speech and offensive language detection in two low-resource Indo-Aryan languages: Sinhala and Gujarati. Sinhala, an official language of Sri Lanka, and Gujarati, a prominent language in India, each pose unique challenges due to their linguistic diversity and limited availability of annotated data. In this paper, these challenges are addressed by employing cutting-edge BERT-based models.

Forum for Information Retrieval Evaluation, December 15-18, 2023, India


*Corresponding authors.

†These authors contributed equally.

✉ gnanasai2111012@ssn.edu.in (G. G. Sai); aswath2111001@ssn.edu.in (A. Venkatesh); kishore2110289@ssn.edu.in (K. N); olirva2110544@ssn.edu.in (O. M); balaji2110065@ssn.edu.in (B. V. A); prabavathyb@ssn.edu.in (P. Balasundaram)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This research paper revolves around the HASOC 2023 shared task [1], which serves as the foundation for our investigation. The shared task includes coarse-grained binary classification, where tweets are categorized into Hate and Offensive (HOF) and Non-Hate and Offensive (NOT) classes. For both languages, participants are provided with a relatively small training set, challenging them to develop effective models for hate speech detection.

This research paper is structured as follows: We begin by presenting a detailed description of the HASOC 2023 task setup for both Sinhala and Gujarati. We then delve into our experimental methodology, which leverages pre-trained BERT models fine-tuned on the provided training data. We explore the transferability of models across languages, aiming to maximize the utility of limited linguistic resources. Finally, we discuss our findings, highlighting the potential impact of our research on mitigating online hate speech and offensive language in these linguistic communities. By combining insights from two distinct languages, our work contributes to a broader understanding of hate speech detection in low-resource language contexts.

2. Related Work

In research [2] conducted by Vinura et al., pre-trained language models for Sinhala text classification were analyzed, with XLM-R being identified as the most effective model. The study introduced high-performing RoBERTa-based monolingual Sinhala models, which offered strong baselines even when there is insufficient labeled data for fine-tuning. It provided valuable usage recommendations and contributed by releasing annotated datasets for future research in Sinhala text classification.

A study [3], conducted by Andrea et al., aimed to assess the suitability of Bidirectional Encoder Representations from Transformers (BERT) models for sentiment analysis and emotion recognition in Twitter data. Two classifiers were developed for each task, and the models were fine-tuned using real-world tweet datasets. The research demonstrated that BERT-based models achieved high accuracy, with 92% for sentiment analysis and 90% for emotion recognition, showcasing BERT's potent language modeling capabilities for text classification in social media data.

The problem of hate speech and offensive language has increased due to the internet being widely used and technical resources being available to most people. If not moderated, it could lead to severe riots and hate-mongering against minorities. Therefore, Apoorv et al. [4], using one vs rest classification, introduced a system to classify a comment as being normal, hateful, or offensive, and the communities targeted by it; a total of 18 labels are used, one for the classification of the comment and the other 17 for the target communities. In addition to the global accuracy, this research study also provided individual accuracy for each community being targeted in the one-vs-rest model.

In the article by Tiwari et al. [5], the focus was on the challenges surrounding hate speech recognition within the realm of social media platforms, with the ultimate goal of enhancing the accuracy of machine learning models. Leveraging Twitter datasets, the researchers conducted a comparative analysis of various machine learning algorithms, considering metrics such as accuracy and precision. Their findings revealed that the combination of XGBoost and TF-IDF embedding yielded the highest accuracy at 94.43%. The article stressed the critical

importance of hate speech detection in ensuring user safety and compliance with laws addressing discriminatory and offensive content.

This study [6] was dedicated to addressing the challenge of identifying and categorizing offensive language and hate speech using cutting-edge text classification techniques. To support the research, the authors curated a custom dataset in the Egyptian Arabic dialect, each manually categorized. They leveraged this dataset to fine-tune and assess various Arabic pre-trained transformer models that employed different transformer architectures and pre-training strategies, specifically tailored for the task of natural language processing and text classification. The results were striking, with an average accuracy of around 96% achieved across all the fine-tuned transformer models, showcasing their efficacy in combating offensive language and hate speech on Egyptian social media platforms.

Ding et al.'s paper [7] introduced an innovative approach known as Hypergraph Attention Networks (HANs) for inductive text classification, with a strong emphasis on efficiency and performance enhancement. HANs leverage hypergraph structures to capture intricate higher-order word relationships within textual data, thereby enriching contextual comprehension. By harnessing sparse hypergraphs, this method effectively curtails computational complexity, rendering it highly scalable, especially for extensive datasets. Experimental outcomes underscore HANs' superiority over existing techniques, showcasing their potential for proficient inductive text classification while efficiently utilizing computational resources.

There was a notable increase in offensive language within the content generated by the crowd across various social platforms. This type of language had the potential to bully or harm the sentiments of individuals or communities. Hajibabae et al [8] had, at that time, delved into investigating and developing various supervised methods and training datasets aimed at automatically detecting or preventing offensive monologues or dialogues. Their experiments yielded promising results in the detection of offensive language using the dataset they had collected from Twitter. After hyperparameter optimization, it was found that three methods—AdaBoost, SVM, and MLP—had achieved the highest average F1-score.

Korde et al.'s paper [9] underscored text mining's commercial potential, as about 80% of data exists in textual form, and unstructured texts offer a rich source of information. The paper centered on text classification, introducing its concept, processing steps, and various classifiers. It conducted a comparative analysis based on criteria like time complexity, principal components, and performance metrics, highlighting text classification's significance in extracting knowledge from unstructured data and aiding classifier selection for diverse applications.

Santhoopa et al.'s paper [10] introduced a deep learning-based approach using a model incorporating Long Short-Term Memory (LSTM) units and FastText word embeddings for hate speech detection. Their model was trained on the Sinhala Unicode Hate Speech dataset from Kaggle, consisting of 6345 Facebook comments, with 54% categorized as hate speech. The study compared this deep learning model with various machine learning algorithms, and the proposed model demonstrated superior performance. It employed pre-trained 100-dimensional FastText word embeddings in its architecture. The model consisted of a Bi-directional LSTM layer, a Dense layer with Rectified Linear Unit (ReLU) activation, and a Sigmoid layer for binary classification. The research suggests the potential for retraining the model for hate speech detection in different languages.

Tanmay et al.'s paper [11] addressed the task of Offensive Language Identification in the

low-resource Indic language Marathi. They focused on a text classification task aimed at discerning offensive from non-offensive tweets. The study evaluated various mono-lingual and multi-lingual BERT models, with a particular focus on those pre-trained with social media data. The performance of MuRIL, MahaTweetBERT, MahaTweetBERT-Hateful, and MahaBERT was compared using the HASOC 2022 test set. Additionally, the paper explored external data augmentation from other existing Marathi hate speech corpora, HASOC 2021 and L3Cube-MahaHate. Notably, MahaTweetBERT, a BERT model pre-trained on Marathi tweets and fine-tuned on a combined dataset (HASOC 2021 + HASOC 2022 + MahaHate), achieved superior performance with an F1 score of 98.43 on the HASOC 2022 test set, establishing a new state-of-the-art result for HASOC 2022 / MOLD v2.

3. Task and Dataset Description

The task focuses on the binary classification of tweets written in Sinhala and Gujarati. The two classes for classification are as follows: 1. Hate and Offensive (HOF): Tweets in this category contain hate speech, profane language, or offensive content targeting individuals or groups based on their characteristics such as race, religion, ethnicity, gender, etc. 2. Non-Hate and Offensive (NOT): Tweets in this category do not contain any hate speech, profanity, or offensive content. They represent neutral or non-harmful expressions in the Sinhala and Gujarati languages. The below subsections discuss Sinhala and Gujarati datasets.

3.1. Sub Task: Identifying Hate, offensive and profane content in Sinhala

The training and test datasets for this task are based on the Sinhala Offensive Language Detection dataset (SOLD) [12]. SOLD is a valuable resource comprising Sinhala tweets labeled for hate speech and offensive content. The dataset is designed to facilitate the development and evaluation of hate speech detection models in the Sinhala language. Participants are encouraged to use this dataset to train their models and subsequently evaluate them on the provided test set.

The task is to classify the tweet as whether it contains hate speech or not. Each entry or row in the CSV file is of the format as given below in Table 1.

Table 1

Attributes of the CSV file for Sinhala dataset

Field	Representation
post_id	represents the unique id of the tweet
text	content of the tweet
label	classification of the tweet

3.2. Sub Task: Identifying Hate, offensive and profane content in Gujarati

The training dataset contains approximately 200 labeled tweets in Gujarati, consisting of both HOF and NOT categories. The evaluation will be conducted on a separate test dataset, ensuring fair evaluation of participant systems' performance.

The task is to classify the tweet as whether it contains hate speech or not. Each entry or row in the CSV file is as per the format given in Table 2.

Table 2

Attributes of the CSV file for Gujarati dataset

Field	Representation
tweet_id	represents the unique id of the tweet
created_at	represents the date when the tweet was posted
text	content of the tweet
user_screen_name	represents the Twitter account name of the tweet creator
label	classification of the tweet

4. Methodologies used

Different NLP architectures like Bert-Based-uncased, Bert-base-multilingual-cased, Sinhala-bert, Gujarathi-bert, and Indic-bert were employed for identifying Hate, offensive, and profane content from tweets in Gujarathi and Sinhala.

4.1. Basic BERT Architecture

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is based on the transformer architecture which relies primarily on attention mechanisms. The BERT model is a multi-layer bidirectional transformer encoder. It consists of an input layer, multiple hidden layers, and an output layer. The input to BERT is a sequence of tokens that are first passed through an embedding layer.

The embedded tokens are then passed to the transformer encoder. The transformer encoder is made up of a stack of identical layers. Each layer consists of two sub-layers - a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The self-attention mechanism allows the model to learn the relationship between different positions in the input sequence to understand the context of the words.

The position-wise feed-forward network applies two linear transformations with a ReLU activation in between each element of the sequence. This helps the model learn complex patterns and relationships between words or tokens from the input. The output of each transformer layer is fed as input to the next layer in the stack. The last hidden state of the first token (which corresponds to the [CLS] token) is used as the aggregate sequence representation for classification tasks.

BERT is trained on two unsupervised prediction tasks - masked language modeling and next-sentence prediction. This allows BERT to learn deep bidirectional representations by conditioning on both left and right contexts in all layers. The pre-trained BERT models can then be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks.

4.2. Bert-Base-uncased

"BERT-base-uncased" is a specific configuration of the BERT model. It shares the same architecture as the original "base BERT" consisting of 12 transformer layers and 110 million parameters. However, the key distinction lies in its vocabulary and tokenization. The "uncased" variant uses an all-lowercase vocabulary, which simplifies tokenization and reduces the vocabulary size compared to the original "base BERT" making it more memory-efficient. This modification allows BERT-base-uncased to be well-suited for NLP tasks that do not require case sensitivity while retaining the strong pre-training and generalization capabilities of the BERT model architecture.

4.3. Bert-base-multilingual-cased

"BERT-base-multilingual-cased" is a specific variant of the BERT model designed for multilingual natural language processing (NLP) tasks. Unlike the original "base BERT" which is primarily trained on English text, this variant is trained on a diverse range of languages. The "cased" aspect indicates that it retains case information in its vocabulary, allowing it to distinguish between uppercase and lowercase letters. This is important for languages where case sensitivity plays a crucial role in understanding context. BERT-base-multilingual-cased is particularly valuable for multilingual applications as it can handle multiple languages effectively, making it a versatile choice for tasks requiring NLP across different linguistic backgrounds.

4.4. Sinhala-bert

"Sinhala BERT" is tailored for the Sinhala language, primarily spoken in Sri Lanka. This specialized variant captures language-specific nuances, script, and context, enabling more effective handling of Sinhala. It proves invaluable for various Sinhala natural language processing tasks, enhancing performance compared to general-purpose BERT models.

4.5. Gujarati-bert

"Gujarati BERT" is designed for the Gujarati language, predominantly spoken in the Indian state of Gujarat. It excels in capturing unique linguistic characteristics, script, and context, making it highly valuable for Gujarati text analysis. Its specialization enhances performance for tasks like text classification, sentiment analysis, and named entity recognition, surpassing general-purpose BERT models.

4.6. Indic-bert

"Indic BERT" is crafted for the diverse family of Indic languages spoken in the Indian subcontinent. Fine-tuned for languages like Hindi, Bengali, Tamil, and more, it adeptly captures linguistic nuances, scripts, and contextual intricacies. Indic BERT proves to be a versatile resource for various natural language processing tasks across the Indic linguistic landscape, outperforming general-purpose BERT models.

5. Result Analysis for Sinhala Dataset

This section discusses the implementation of various NLP techniques for Sinhala and Gujarati Datasets with the analysis of the results using evaluation metrics namely Macro-F1, Macro-Precision, and Macro-Recall.

5.1. Implementation

The datasets for Sinhala and Gujarati text classification share a similar structure, each containing specific columns such as "post_id", "tweets", and "labels", and "tweet_id", "created_at", "text", "user_screen_name", and "label" respectively. Data preparation involves designating the "labels" or "label" column as the target variable, with the corresponding "tweets" or "text" column housing the data to be categorized as offensive or not. Both datasets are then divided into an 80% training set and a 20% testing set, facilitating the assessment of model performance for hate speech detection in these languages.

For the task of classifying Sinhala offensive tweets, five BERT-based models denoted as **M1**, **M2**, **M3**, **M4**, and **M5** have been selected. These models encompass various BERT architectures fine-tuned for different languages and tasks. Specifically, **M1**, **M2**, **M3**, **M4**, and **M5** correspond to Bert-base-cased, Bert-base-multilingual-cased, Sinhala-bert, Bert-base-multilingual-uncased, and Indic-bert, respectively. Each model's respective tokenizer is applied to convert tweet text into suitable numerical representations for BERT-based analysis.

The models **M1**, **M2**, **M3**, **M4**, and **M5** are employed for classifying Sinhala offensive text, each with specific attributes. **M1**, utilizing "bert-base-cased", primarily designed for English but capable of handling Sinhala, employs WordPiece tokenization for subword units, considering frequently occurring Sinhala character sequences. **M2**, known as "bert-base-multilingual-cased", is a multilingual model, effectively tokenizing Sinhala text for multilingual tasks with a broader vocabulary. **M3**, "Sinhala-bert", is tailored specifically for Sinhala, employing a Sinhala-specific tokenizer for subword tokenization. **M4**, "bert-base-multilingual-uncased", like **M2**, handles subword tokenization but without case differentiation, suitable for processing Sinhala and other languages. Lastly, **M5**, "Indic-bert," designed for the Indian subcontinent, including Sinhala, uses a subword tokenization technique customized for Indic scripts, optimizing its ability to handle Sinhala-specific character combinations, linguistic patterns, and word segmentation.

For offensive Gujarati tweet classification, five BERT-based models denoted as **N1**, **N2**, **N3**, **N4**, and **N5** are chosen, and fine-tuned for various languages and tasks. These models correspond to bert-base-multilingual-uncased, Indic-bert, bert-base-multilingual-cased, Gujarati-bert, and

Gujarati-bert (with preprocessing), respectively. Each model's specific tokenizer is applied to convert tweet text into suitable numerical representations for BERT analysis.

The models, **N1**, **N2**, **N3**, **N4**, and **N5**, are utilized for the classification of Gujarati offensive text in a multilingual context. **N1**, employing "bert-base-multilingual-uncased", is designed to handle various languages, including Gujarati, utilizing subword tokenization for text segmentation. **N2**, known as "Indic-bert", is specialized for Indian subcontinent languages, with a tailored subword tokenization technique to accurately represent Gujarati text. **N3**, "bert-base-multilingual-cased", is a multilingual model that can handle Gujarati and employs subword tokenization similar to **N1** but with a broader vocabulary. **N4**, "Gujarati-bert", is customized for the Gujarati language, featuring a dedicated tokenizer and subword tokenization optimized for Gujarati script characteristics. **N5**, "Gujarati-bert(with preprocessing)", is an enhanced version of "Gujarati-bert", incorporating preprocessing steps to improve performance on Gujarati text. These preprocessing steps encompass text cleaning, normalization, and other language-specific enhancements.

These tokenized inputs are then fed into the models for training and testing. The training process involves specifying hyperparameters like batch size, the number of training epochs, and learning rate. Additionally, appropriate optimization algorithms, such as AdamW, are employed, along with learning rate schedulers to fine-tune the models.

After the training phase, the models are evaluated on separate test datasets, each specific to its respective language. The Sinhala dataset consists of 1500 rows, maintaining column names such as post_id, tweets, and labels. This testing phase assesses each model's ability to generalize and make accurate predictions on new, unseen Sinhala text. Similarly, the Gujarati dataset contains 40 rows with columns like tweet_id, text, and label. The testing phase for Gujarati evaluates each model's performance in making accurate predictions on new and unseen Gujarati tweets, ensuring a comprehensive assessment of their capabilities in both language contexts.

The complete implementation of these models, along with the code, can be found on our GitHub repository at https://github.com/g-sai/HASOC2023-SSN_CSE_ML_TEAM.

5.2. Results and discussion

The models **M1**, **M2**, **M3**, **M4**, and **M5** were applied to classify text data for hate speech detection in Sinhala. To evaluate their performance, we employed evaluation metrics, including Macro-F1, Macro-Precision, and Macro-Recall. These metrics allow us to assess the model's abilities to accurately identify and predict instances of hate speech within Sinhala text. Upon analyzing the results in Table 3, it becomes evident that **M3** achieved the best results among all the models, with an impressive Macro-F1 score of 0.7946. This score highlights its proficiency in correctly identifying hate speech within the Sinhala text data, outperforming the other models and indicating superior performance in hate speech detection.

Similarly, models **N1**, **N2**, **N3**, **N4**, and **N5** were utilized to classify the text data for hate speech detection in Gujarati. We computed evaluation metrics, including Macro-F1, Macro-Precision, and Macro-Recall, to gauge their performance. After scrutinizing the results in Table 4, it's clear that **N5** yielded the best results among all the models, achieving an impressive Macro-F1 score of 0.7732. This score underscores its proficiency in correctly identifying hate speech within Gujarati text data, outperforming the other models and indicating superior performance in hate

speech detection.

Table 3

Assessment of Models using Evaluation Metrics

Model	Macro-F1	Macro-Precision	Macro-Recall
Bert-Base-cased(M1)	0.4815	0.5251	0.5144
Bert-base-multilingual-cased (M2)	0.7890	0.7869	0.7931
Sinhala Bert(M3)	0.7946	0.7977	0.7923
Bert-base-multilingual-uncased(M4)	0.3156	0.3229	0.3344
Indic Bert(M5)	0.7936	0.7914	0.7980

Table 4

Assessment of Models using Evaluation Metrics

Model	Macro-F1	Macro-Precision	Macro-Recall
Bert-base-multilingual-uncased(N1)	0.6316	0.6522	0.6759
Indic Bert(N2)	0.2598	0.2826	0.2522
Bert-base-multilingual-cased(N3)	0.6446	0.6416	0.6500
Gujarati Bert(N4)	0.7558	0.7563	0.7959
Gujarati Bert(with preprocessing) (N5)	0.7732	0.7676	0.8048

6. Conclusion

In this study, we evaluated the effectiveness of several BERT-based models for detecting hate speech and offensive language in Sinhala and Gujarati tweets. Our results demonstrated that the Sinhala-BERT model achieved the highest Macro-F1 score of 0.7946 for identifying hateful content in Sinhala tweets. For the Gujarati data, the Gujarati-BERT model with preprocessing steps showed the best performance with a Macro-F1 score of 0.7732. Overall, the findings suggest that language-specific models that leverage characteristics of the target script can more accurately classify instances of hateful and offensive content compared to general-purpose multilingual models. This research contributes to developing automated techniques for moderating social media in under-resourced languages like Sinhala and Gujarati while promoting inclusive online discussions. The models and datasets introduced in this study can also serve as valuable resources for future NLP research on these languages.

7. References

- [1] Shrey Satapara, Hiren Madhu, Tharindu Ranasinghe, Alphaeus Eric Dmonte, Marcos Zampieri, Pavan Pandya, Nisarg Shah, Modha Sandip, Prasenjit Majumder, and Thomas

- Mandl, *Overview of the HASOC Subtrack at FIRE 2023: Hate-Speech Identification in Sinhala and Gujarati*, In Kripabandhu Ghosh, Thomas Mandl, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India, December 15-18, 2023*, CEUR Workshop Proceedings, CEUR-WS.org, 2023.
- [2] Chiorrini, Andrea & Diamantini, Claudia & Mircoli, Alex & Potena, Domenico. "Emotion and sentiment analysis of tweets using BERT." March 2021.
 - [3] Dhananjaya, Vinura & Demotte, Piyumal & Ranathunga, Surangika & Jayasena, Sanath. "BERTifying Sinhala – A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification." *10.48550/arXiv.2208.07864*, August 2022.
 - [4] A. Aditya, R. Vinod, A. Kumar, I. Bhowmik and J. Swaminathan, "Classifying Speech into Offensive and Hate Categories along with Targeted Communities using Machine Learning," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 291-295, doi: 10.1109/ICICT54344.2022.9850944.
 - [5] Tiwari, Abhay, and Anupam Agrawal. "Comparative Analysis of Different Machine Learning Methods for Hate Speech Recognition in Twitter Text Data." In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICT), pp. 1016-1020. IEEE, 2022.
 - [6] I. Ahmed, M. Abbas, R. Hatem, A. Ihab and M. W. Fahkr, "Fine-tuning Arabic Pre-Trained Transformer Models for Egyptian-Arabic Dialect Offensive Language and Hate Speech Detection and Classification," 2022 20th International Conference on Language Engineering (ESOLEC), Cairo, Egypt, 2022, pp. 170-174, doi: 10.1109/ESOLEC54569.2022.10009167.
 - [7] Ding, Kaize, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. "Be more with less: Hypergraph attention networks for inductive text classification." *arXiv preprint arXiv:2011.00387* (2020).
 - [8] Parisa Hajibabae, Masoud Malekzadeh, Mohsen Ahmadi, Maryam Heidari, Armin Esmaeilzadeh, Reyhaneh Abdolazimi, James H Jr Jones, "Offensive Language Detection on Social Media Based on Text Classification," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0092-0098, doi: 10.1109/CCWC54503.2022.9720804.
 - [9] Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3, no. 2 (2012): 85.
 - [10] Fernando, W. S. S., Ruvan Weerasinghe, and E. R. A. D. Bandara. "Sinhala hate speech detection in social media using machine learning and deep learning." In 2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 166-171. IEEE, 2022.
 - [11] Chavan, Tanmay, Shantanu Patankar, Aditya Kane, Omkar Gokhale, and Raviraj Joshi. "A Twitter BERT Approach for Offensive Language Detection in Marathi." *arXiv preprint arXiv:2212.10039* (2022).
 - [12] Tharindu Ranasinghe, Ishara Anuradha, Danushka Premasiri, Kasun Silva, Hirunima Hettiarachchi, Lakshika Uyangodage, and Marcos Zampieri. "SOLD: Sinhala Offensive Language Dataset." *arXiv preprint arXiv:2212.00851*, 2022.