Sinhala and Gujarati Hate Speech Detection

M. Krithik Sathya[1], K.H. Gopalakrishnan[1], Manickam PA[1] and Prabavathy Balasundaram[2]

[1]*UG Student, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India*

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

Abstract

This study, conducted by the ”Krispy Mango” research team, focuses on hate speech and offensive content detection in two low-resource Indo-Aryan languages, Sinhala and Gujarati, as part of the HASOC 2023 shared tasks. We address the difficulty of classifying tweets into Hate and Offensive (HOF) and Non-Hate and Offensive (NOT) categories by fine-tuning the BERT models. This work presents findings in the form of macro F1 scores and precision metrics for both languages. Our approach aims to advance the state-of-the-art in detecting hate speech while taking into account the particular linguistic characteristics and resource restrictions of these languages.

Keywords

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

1. Introduction

The digital age has revolutionized the way we communicate and connect with one another, primarily through the widespread adoption of social media platforms. However, this unprecedented level of global interconnectivity has also brought about a concerning surge in hate speech and offensive content. Effectively addressing this challenge and developing robust methods for detecting and countering hate speech has become imperative. This research aims to contribute significantly to this effort by focusing on hate speech detection in two South Asian languages, Sinhala and Gujarati.

While hate speech detection in English has received substantial attention, these languages have received comparatively less consideration in the realm of Natural Language Processing (NLP). Hate speech is a pervasive issue that crosses linguistic boundaries, emphasizing the importance of developing models that can identify such content in non-English languages as effectively as in English. Sinhala, spoken in Sri Lanka, and Gujarati, a major Indian language, pose unique linguistic challenges due to their complex structures and distinct scripts. Detecting hate speech in these languages demands tailored approaches and models capable of understanding their intricacies. [1]

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

krithik2110693@ssn.edu.in (M. K. Sathya); gopalakrishnan2110375@ssn.edu.in (K.H. Gopalakrishnan); manickam2110305@ssn.edu.in (M. PA); prabavathyb@ssn.edu.in (P. Balasundaram)

0000-0002-0877-7063 (M. K. Sathya); 0000-0001-7116-9338 (K.H. Gopalakrishnan); 0000-0002-9421-8566 (M. PA); 0000-0002-9421-8567 (P. Balasundaram)

The following is the structure of this research paper: We start by providing a thorough explanation of the Sinhala and Gujarati task configuration for HASOC 2023. We next dive into our experimental methodology, which employs pre-trained BERT models fine-tuned on the available training data. In order to make the most of the restricted linguistic resources, we investigate the transferability of models across linguistic boundaries. Finally, we highlight our research's possible effects on reducing online hate speech and harmful language in different linguistic communities as we examine our findings. Our work advances knowledge of hate speech identification in low-resource language circumstances by merging ideas from two different languages. [2] [3]

2. Related Works

In recent research by Vinura et al. [4], the effectiveness of pre-trained language models for Sinhala text classification was explored. Among these models, XLM-R emerged as the most potent choice. The study introduced RoBERTa-based monolingual Sinhala models, establishing strong baselines, even in the presence of limited labelled data. Additionally, this research made significant contributions by releasing annotated datasets, providing valuable resources for future studies in Sinhala text classification.

Andrea et al. [5] conducted a study focusing on the applicability of Bidirectional Encoder Representations from Transformers (BERT) models for sentiment analysis and emotion recognition in Twitter data. Through the development and fine-tuning of two classifiers for each task, they achieved remarkable results, with BERT-based models achieving accuracy rates of 92 per cent for sentiment analysis and 90 per cent for emotion recognition. These findings underscored BERT's proficiency in modelling language for text classification within the realm of social media data.

Tiwari et al. [6] directed their efforts towards addressing challenges in hate speech recognition within the context of social media platforms. They conducted a comparative analysis of various machine learning algorithms, emphasizing accuracy and precision metrics. Their findings identified the combination of XGBoost and TF-IDF embedding as the highest-performing approach, achieving an accuracy rate of 94.43 per cent. This research emphasized the critical role of hate speech detection in promoting user safety and compliance with laws addressing offensive content.

Wang et al. [7] offered a comprehensive retrospective on the evolution of text classification, spanning traditional shallow learning techniques to deep learning models. Their meticulous examination of six pivotal methods, including ReNN, MLP, RNN, CNN, Attention, and Transformer, highlighted their respective strengths and limitations. The paper underscored the dominance of deep learning models in text classification and highlighted ongoing research in attention mechanisms, Transformers, robustness, and graph neural networks, indicating the continuous evolution of text classification solutions.

Ding et al. [7] introduced an innovative approach, Hypergraph Attention Networks (HANs), for inductive text classification. With a focus on efficiency and performance enhancement, HANs harnessed hypergraph structures to capture intricate word relationships within textual data. By utilizing sparse hypergraphs, this method effectively managed computational complexity, showcasing its scalability for extensive datasets. Experimental results underscored HANs' superiority over existing techniques, demonstrating their potential for proficient inductive text classification while efficiently utilizing computational resources.

Minaee et al. [8] conducted an extensive review comparing deep learning models to classical machine learning in text classification tasks such as sentiment analysis and news categorization. They evaluated over 150 recent deep learning-based text classification models, providing insights into technical innovations and strengths. The paper also analyzed the performance of these models on benchmark datasets, supporting their effectiveness with empirical evidence. It concluded by outlining potential avenues for future research, serving as a valuable resource for understanding the current landscape and future potential of deep learning in text classification.

3. Task and Dataset Description

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

The task focuses on categorizing tweets published in Sinhala in a binary form. The two classification categories are as follows: 1. Hate and unpleasant (HOF): Tweets that target people or groups based on attributes like race, religion, ethnicity, gender, etc. are included in this category. They may also use profanity or other unpleasant language. 2. Non-Hate and Offensive (NOT): Tweets falling under this category do not contain any offensive language, profanity, or hate speech. They represent neutral or non-harmful expressions in the Sinhala language. The train/ test sets are based on the recently released SOLD: Sinhala Offensive Language Detection dataset. [9]

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 task focuses on categorizing tweets published in Gujarati in a binary form. The two classification categories are as follows: 1. Hate and unpleasant (HOF): Tweets that target people or groups based on attributes like race, religion, ethnicity, gender, etc. are included in this category. They may also use profanity or other unpleasant language. 2. Non-Hate and Offensive (NOT): Tweets falling under this category do not contain any offensive language, profanity, or hate speech. They represent neutral or non-harmful expressions in the Gujarati language.

Table 2
Attributes of the CSV file for Gujarati dataset

Field	Representation
post id	Represents the unique id of the tweet
text	Content of the tweet
label	Classification of the tweet

4. Methodologies used

Different NLP architectures like xlm-roberta-base, bert-base-multilingual-cased, intfloat/multilingual-e5-base, openai/whisper-large, Sinhala bert, Gujarathi-bert, were employed for identifying Hate, offensive, and profane content from tweets in Gujarathi and Sinhala.

4.1. Basic BERT Architecture

The BERT model, an acronym for ”Bidirectional Encoder Representations from Transformers,” is grounded in the transformer architecture, emphasizing attention mechanisms. Comprising a multi-layer bidirectional transformer encoder, it includes an input layer, multiple hidden layers, and an output layer. Input sequences undergo initial processing through an embedding layer before entering the transformer encoder.[10]

This encoder consists of a stack of uniform layers, each housing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The self-attention mechanism enables the model to discern interrelations among input sequence positions, aiding contextual comprehension.

The position-wise feed-forward network applies two linear transformations, with ReLU activations interleaved, to each sequence element, enabling the model to capture intricate patterns and interdependencies among input tokens. Importantly, the final hidden state of the initial token ([CLS]) serves as the holistic sequence representation for classification tasks.

BERT undergoes training through two unsupervised prediction tasks: masked language modeling and next-sentence prediction. This dual training equips BERT with profound bidirectional representations, leveraging contextual information from both preceding and subsequent contexts across all layers. Pre-trained BERT models can then be fine-tuned with an additional output layer, making them adaptable and potent tools for diverse natural language processing (NLP) tasks.[11][12]

4.2. XLM-RoBERTa

The ”XLM-RoBERTa” model represents a powerful fusion of two renowned architectures: XLM (Cross-lingual Language Model) and RoBERTa. This variant excels in multilingual natural language processing tasks, with an emphasis on cross-lingual understanding. It boasts a vast parameter count and a deep architecture comprising multiple transformer layers. XLM-RoBERTa is pre-trained on an extensive corpus encompassing a multitude of languages, allowing it to comprehend and generate text in a wide array of linguistic contexts. Notably, it does not differentiate between uppercase and lowercase letters, ensuring robust performance in both case-sensitive and case-insensitive scenarios. This model's versatility and cross-lingual capabilities make it an invaluable asset for researchers and practitioners engaged in multilingual NLP tasks, ranging from machine translation to document classification. [4]

4.3. Bert-base-mutilingual-cased

"BERT-base-multilingual-cased" is a BERT model version designed for multilingual natural language processing (NLP) applications. Unlike the original "base BERT," which was trained exclusively on English text, this variation was trained on a variety of languages. The model has 6 layers, 768 dimensions and 12 heads, totalizing 134M parameters (compared to 177M parameters for mBERT-base). On average, this model, referred to as DistilmBERT, is twice as fast as mBERT-base. The "cased" element denotes that it stores case information in its lexicon, allowing it to differentiate between uppercase and lowercase letters. This is critical for languages where case sensitivity is critical for interpreting context. BERT-base-multilingual-cased is very useful for multilingual applications since it can efficiently handle many languages, giving it a versatile solution for tasks needing NLP across different linguistic backgrounds.

4.4. intfloat/multilingual-e5-base

"intfloat/multilingual-e5-base" is a specialized BERT variant developed to address the demands of multilingual natural language processing. It offers a comprehensive solution for tasks involving diverse languages and linguistic characteristics. Trained on an extensive multilingual corpus, this model leverages transformer-based architecture and deep neural networks to facilitate effective language understanding and generation. Notably, it encompasses a cased vocabulary, enabling it to preserve case information, which is pivotal in languages where case sensitivity plays a significant role in semantic interpretation. This variant's adaptability and multilingual competence render it a valuable tool for cross-lingual applications such as multilingual document classification, sentiment analysis, and more.

4.5. OpenAI/Whisper

"OpenAI/Whisper-Large" is a large-scale automatic speech recognition (ASR) model designed to transcribe spoken language into text. This model's capabilities are underpinned by a massive architecture, extensive pre-training on diverse audio data, and a robust transformer-based architecture. It excels in recognizing speech across multiple languages and dialects, making it a versatile choice for ASR tasks in various linguistic contexts. With its remarkable capacity for handling large volumes of spoken data and its ability to adapt to distinct accents and acoustic conditions, OpenAI/Whisper-Large is a valuable asset for applications such as transcription services, voice assistants, and more, where accurate speech-to-text conversion is paramount.

4.6. keshan/SinhalaBERTo

Keshan/SinhalaBERTo is a specialized language model developed to address the unique challenges posed by the Sinhala language. Sinhala, being a low-resource language, has limited access to pre-trained language models. Keshan/SinhalaBERTo fills this gap as a slightly smaller but highly valuable language model. It is trained on the OSCAR Sinhala dedup dataset, making it a relevant resource for Sinhala natural language processing tasks. The model specifications, including a vocabulary size of 52,000, max position embeddings of 514, 12 attention heads, 6 hidden layers, and a type vocabulary size of 1, create a robust foundation for Keshan/SinhalaBERTo, a specialized language model for Sinhala text processing. [4]

4.7. Gujarati-bert

”Gujarati BERT” is a modified variant of the BERT model built exclusively for the Gujarati language, which is widely spoken in the Indian state of Gujarat and other areas. Gujarati BERT is fine-tuned for Gujarati text, as opposed to the usual ”base BERT” model, which is trained on a varied variety of languages. This allows it to capture the specific linguistic qualities, script, and context of the Gujarati language more efficiently. Gujarati BERT is particularly useful for natural language processing tasks in Gujarati, such as text categorization, sentiment analysis, and named entity recognition, due to this speciality. When compared to the more general-purpose base BERT model, Gujarati BERT’s domain expertise improves its performance and applicability in the context of the Gujarati language.

5. Result Analysis for Sinhala Dataset

5.1. Implementation

In this section, we present the results of our offensive tweet classification task, employing five diverse BERT-based models: M1 (XLM-RoBERTa), M2 (Keshan/SinhalaBERTo), M3 (Bert-base-multilingual-cased), M4 (Bert-base-multilingual-uncased), and M5 (intfloat/multilingual-e5-base). These models have distinct linguistic characteristics and tokenization methods, which contribute to their unique performance.

Model M1 is based on XLM-RoBERTa, exhibits robust performance in classifying offensive tweets. XLM-RoBERTa’s multilingual competence allows it to handle a wide range of languages effectively, including Sinhala. Its tokenization strategy considers various linguistic nuances, and it demonstrates a strong ability to generalize across languages. This model’s adaptability and pre-training on diverse multilingual data contribute to its high classification accuracy on the test dataset.

M2 is powered by Keshan/SinhalaBERTo, which is tailored explicitly for the Sinhala language. Its tokenizer, optimized for Sinhala text, excels in capturing the language’s unique characteristics. This model showcases impressive results in classifying offensive tweets, demonstrating the importance of language-specific models in achieving high accuracy on Sinhala text. M2’s fine-tuning on Sinhala data contributes to its superior Sinhala text understanding and classification capabilities.

Model M3 is Bert-base-multilingual-cased, it is designed as a versatile, multilingual BERT variant. Although not optimized exclusively for Sinhala, it manages to handle Sinhala text effectively due to its extensive multilingual vocabulary. M3’s tokenization, which is akin to bert-base-cased, successfully translates Sinhala text into subword tokens, allowing it to perform well in cross-lingual offensive tweet classification.

M4 is Bert-base-multilingual-uncased, this shares similarities with M3 but lacks case sensitivity. Despite this difference, it effectively tokenizes Sinhala text, thanks to its subword tokenization method and multilingual vocabulary. M4 showcases commendable performance in the classification task, affirming its suitability for processing Sinhala and other languages without consideration for letter casing.

Model M5 is intfloat/multilingual-e5-base, it is geared towards multilingual natural language processing tasks. Its subword tokenization and extensive pre-training enable it to handle Sinhala text with competence. M5 exhibits competitive results in classifying offensive tweets, highlighting its adaptability and cross-lingual proficiency.

These tokenized inputs are then used to train and test the models. During the training phase, hyperparameters such as batch size, number of training epochs, and learning rate must be specified. To fine-tune the models, appropriate optimization algorithms, such as AdamW, are used in conjunction with learning rate schedulers. Following the training phase, the models are tested on a separate 1500-row test dataset with the same column names as the training data (post id, tweets, labels). During the testing phase, each model's capacity to generalize and generate correct predictions on new, unseen data is evaluated. [13]

5.2. Results and discussion

To categorize text data for hate speech detection in Sinhala, the models M1, M2, M3, M4, and M5 were used. To examine the performance of these models, evaluation metrics such as Macro-F1, Macro-Precision, and Macro Recall were generated. These metrics provide insight on the model's ability to reliably identify and predict instances of hate speech in Sinhala text data. After examining the findings for these assessment measures in Table 3, it is clear that M5 outperforms all other models.

Macro-F1 was used to assess these models because it combines precision and recall into a single score, offering a balanced estimate of the model's capacity to reliably categorize instances of hate speech. M5, with a stellar Macro-F1 score of 0.8371, demonstrated its proficiency in correctly identifying hate speech within the Sinhala text data, outperforming the other models. A higher Macro-F1 score suggests superior performance in hate speech detection.

Table 3
Assessment of Models using Evaluation Metrics

Model	Macro-F1	Macro-Precision	Macro-Recall
XLM-RoBERTa(M1)	0.7210	0.6803	0.7604
Keshan/SinhalaBERTo(M2)	0.6451	0.6532	0.6430
Bert-base-multilingual-cased(M3)	0.8141	0.8125	0.8162
Bert-base-multilingual-uncased(M4)	0.7728	0.7813	0.7776
intfloat/multilingual-e5-base(M5)	0.8371	0.8439	0.8326

6. Result Analysis for Gujarati

6.1. Implementation

In this section, we present the results of our offensive tweet classification task for the Gujarati language, utilizing five distinct BERT-based models: M1 (XLM-RoBERTa), M2 (bert-base-multilingual-cased), M3 (bert-base-multilingual-uncased), M4 (OpenAI/Whisper-Large), and M5 (Gujarati BERT). These models vary in terms of their linguistic capabilities and tokenization methods, which influence their performance on the Gujarati dataset.

Model M1 is based on XLM-RoBERTa, it demonstrates strong performance in classifying offensive tweets in Gujarati. XLM-RoBERTa's multilingual capabilities allow it to handle a wide range of languages, including Gujarati, effectively. Its tokenization strategy considers linguistic nuances, and the model exhibits a robust ability to generalize across languages. M1's adaptability and pre-training on diverse multilingual data contribute to its high classification accuracy on the test dataset.

M2 which utilises bert-base-multilingual-cased, is a versatile, multilingual BERT variant. Although not optimized exclusively for Gujarati, it effectively handles Gujarati text due to its extensive multilingual vocabulary. M2's tokenization method successfully translates Gujarati text into subword tokens, enabling it to perform well in cross-lingual offensive tweet classification.

Model M3 is bert-base-multilingual-uncased, which shares similarities with M2 but lacks case sensitivity. Despite this difference, it effectively tokenizes Gujarati text, thanks to its subword tokenization method and multilingual vocabulary. M3 showcases commendable performance in the classification task, affirming its suitability for processing Gujarati and other languages without consideration for letter casing.

M4 is powered by OpenAI/Whisper-Large, it is designed for large-scale automatic speech recognition (ASR). While not specifically tailored for text classification, its robust architecture allows it to capture spoken Gujarati language effectively. This model showcases competitive results in the offensive tweet classification task, demonstrating its adaptability beyond ASR, especially in tasks involving Gujarati text.

Model M5 is Gujarati BERT, it is a specialized variant designed explicitly for the Gujarati language. Its tokenizer is tailored to handle Gujarati text's unique characteristics effectively. M5 demonstrates impressive results in classifying offensive tweets, emphasizing the importance of language-specific models in achieving high accuracy on Gujarati text. Its fine-tuning on Gujarati data contributes to its superior Gujarati text understanding and classification capabilities.[14]

These tokenized inputs are then used to train and test the models. During the training phase, hyperparameters such as batch size, number of training epochs, and learning rate must be specified. To fine-tune the models, appropriate optimization algorithms, such as AdamW, are used in conjunction with learning rate schedulers. Following the training phase, the models are tested on a separate 1500-row test dataset with the same column names as the training data (post id, tweets, labels). During the testing phase, each model's capacity to generalize and generate correct predictions on new, unseen data is evaluated.

6.2. Results and discussion

To categorize text data for hate speech detection in Sinhala, the models M1, M2, M3, M4, and M5 were used. To examine the performance of these models, evaluation metrics such as Macro-F1, Macro-Precision, and Macro Recall were generated. These metrics provide insight on the model's ability to reliably identify and predict instances of hate speech in Sinhala text data. After examining the findings for these assessment measures in Table 3, it is clear that M2 outperforms all other models.

Macro-F1 was used to assess these models because it combines precision and recall into a single score, offering a balanced estimate of the model's capacity to reliably categorize instances of hate speech. M2, with a stellar Macro-F1 score of 0.7956, demonstrated its proficiency in correctly identifying hate speech within the Gujarati text data, outperforming the other models. A higher Macro-F1 score suggests superior performance in hate speech detection.

Table 4
Assessment of Models using Evaluation Metrics

Model	Macro-F1	Macro-Precision	Macro-Recall
XLM-RoBERTa(M1)	0.7210	0.6803	0.7604
Bert-base-multilingual-cased(M2)	0.7956	0.7897	0.8035
Bert-base-multilingual-uncased(M3)	0.7739	0.7658	0.7955
OpenAI/Whisper-Large(M4)	0.7409	0.7453	0.7386
Gujarati BERT(M5)	0.6415	0.6609	0.6861

7. Conclusion

In this study, we evaluated the efficacy of multiple BERT-based models for detecting hate speech and abusive language in Sinhala and Gujarati tweets. We tested the efficacy of multiple BERT-based models for detecting hate speech and abusive language in Sinhala and Gujarati tweets in this study. The intfloat/multilingual-e5-base model earned the highest Macro-F1 score of 0.8371 for detecting hateful content in Sinhala tweets. The bert-base-multilingual-cased model with preprocessing steps performed best for the Gujarati data, with a Macro-F1 score of 0.7956.

Overall, the findings suggest that multilingual models outperform low-resource, language-specific models in terms of F1 scores. This performance advantage can be attributed to their access to a larger and more diverse dataset. Multilingual models are trained on text from a wide range of languages, which inherently provides a richer linguistic context and a broader spectrum of language patterns. This diversity allows them to capture cross-lingual insights and generalize better across various languages, including low-resource ones. In contrast, low-resource language-specific models, with limited training data, struggle to grasp the full language complexity. Their effectiveness is hindered by data scarcity, limiting their ability to adapt to nuances and context. Higher F1 scores of multilingual models emphasize the advantage of diverse training data. This highlights the significance of data availability, especially for low-resource languages. It underscores the potential for further advancements to enhance language-specific models in the future.

This study adds to the development of automated approaches for moderating social media in underserved languages such as Sinhala and Gujarati, while also encouraging inclusive online debates. The models and datasets presented in this paper can also serve as valuable resources for future NLP research on these languages.

References

[1] B. Di Fátima, Hate speech on social media: A global approach (2023). doi:10.25768/654-916-9.

[2] S. Satapara, H. Madhu, T. Ranasinghe, A. E. Dmonte, M. Zampieri, P. Pandya, N. Shah, M. Sandip, P. Majumder, T. Mandl, Overview of the hasoc subtrack at fire 2023: Hate-speech identification in sinhala and gujarati, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[3] B. R. Chakravarthi, B. Bharathi, C. O'Riordan, H. Murthy, T. Durairaj, T. Mandl, et al., Speech and Language Technologies for Low-Resource Languages: First Inter-national Conference, SPELLL 2022, Kalavakkam, India, November 23–25, 2022, Pro-ceedings, Springer Nature, 2023. doi:10.1007/978-3-031-33231-9.

[4] V. Dhananjaya, P. Demotte, S. Ranathunga, S. Jayasena, Bertifying sinhala–a comprehensive analysis of pre-trained language models for sinhala text classification, arXiv preprint arXiv:2208.07864 (2022).

[5] A. Chiorrini, C. Diamantini, A. Mircoli, D. Potena, Emotion and sentiment analysis of tweets using bert., in: EDBT/ICDT Workshops, volume 3, 2021.

[6] A. Tiwari, A. 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 (ICICICT), IEEE, 2022, pp. 1016–1020.

[7] K. Ding, J. Wang, J. Li, D. Li, H. Liu, Be more with less: Hypergraph attention networks for inductive text classification, arXiv preprint arXiv:2011.00387 (2020).

[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: a comprehensive review, ACM computing surveys (CSUR) 54 (2021) 1–40.

[9] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, arXiv preprint arXiv:2212.00851 (2022).

[10] Z. Wang, Deep learning based text classification methods, Highlights in Science, Engineering and Technology 34 (2023) 238–243.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] V. Korde, C. N. Mahender, Text classification and classifiers: A survey, International Journal of Artificial Intelligence & Applications 3 (2012) 85.

[13] W. Fernando, R. Weerasinghe, E. 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), IEEE, 2022, pp. 166–171.

[14] T. Chavan, S. Patankar, A. Kane, O. Gokhale, R. Joshi, A twitter bert approach for offensive language detection in marathi, arXiv preprint arXiv:2212.10039 (2022).