

# Breaking Barriers: Multilingual Toxicity Analysis for Hate Speech and Offensive Language in Low-Resource Indo-Aryan Languages

Ch Muhammad Awais<sup>1</sup>, Jayveersinh Raj<sup>2</sup>

<sup>1</sup>University of Pisa, Pisa, Italy

<sup>2</sup>Innopolis University, Innopolis, Russia

## Abstract

In our interconnected digital landscape, tackling the proliferation of hate speech and offensive content across languages is a pressing challenge. This paper presents a pioneering approach, fine-tuning the XLM-RoBERTa model for hate speech detection in low-resource languages, transcending linguistic boundaries. Our work is motivated by our participation in the Hate Speech and Offensive Content Identification (HASOC) competition, where our team, AI Alchemists, achieved 3rd place in Sinhala, and participated in Task 1 and 4 with the provided datasets. Our average position was 4th, with no position less than 6th.

Our experiments show that the model can be used to detect hate speech in many different languages, including Sinhala and Bodo, with high accuracy (F1 scores of 0.8345 and 0.844, respectively). It also achieved promising results on Gujarati, Bengali, and Assamese (F1 scores of 0.793, 0.726, and 0.707, respectively).

We emphasize that the quality and size of the training dataset are important factors in the performance of the model. With further research and access to more balanced datasets, we believe that our model has the potential to outperform state-of-the-art models and curb hate speech across diverse linguistic landscapes, fostering more inclusive online spaces.

## Keywords

Multilingual Toxicity Analysis, Hate Speech Detection, Low-Resource Languages, Natural Language Processing, XLM-RoBERTa Model, Transfer Learning, Zero-Shot Transfer, Offensive Content, Cross-Lingual Adaptability, Linguistic Barriers

## 1. Introduction

In today's digital age, the widespread proliferation of offensive content has become an urgent societal concern. Hate speech, offensive language, and profanity have found fertile ground on various online platforms, perpetuating hostility and division among users [1]. Effectively addressing this issue requires advanced tools and methodologies capable of identifying and mitigating offensive content with precision since it is challenging to categorize tweets or comments without obvious hateful keyword [2], and partly because there isn't a consensus definition of hate speech, examination of its demographic influences, or investigation of its most potent

---


*Forum for Information Retrieval Evaluation, December 9-13, 2023, India*

✉ c.awais@studenti.unipi.it (C. M. Awais); j.raj@innopolis.university (J. Raj)

🌐 <https://cm-awais.github.io/> (C. M. Awais); <https://jayveersinh-raj.github.io/> (J. Raj)



© 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)

characteristics [3]. Natural Language Processing (NLP) has emerged as a powerful approach in this endeavor, harnessing the fusion of machine learning and language analysis to discern and counteract toxic language [4].

NLP techniques have proven instrumental in analyzing vast volumes of textual data, enabling the detection of offensive content across diverse linguistic landscapes. By leveraging computational linguistics and statistical models, NLP algorithms can automatically classify texts into offensive and non-offensive categories [5]. The ability to aggregate feedback without human intervention makes this incredibly helpful [6]. These algorithms excel at recognizing patterns, contextual cues, and linguistic features associated with offensive language, thereby aiding in the identification and moderation of problematic content [7, 8].

However, while NLP has demonstrated remarkable promise in addressing offensive content, its application to Low-Resource Indo-Aryan Languages presents a unique set of challenges. Low-resource languages, such as Sinhala and Gujarati, often lack the extensive datasets and linguistic resources available for major languages like English. Consequently, developing effective NLP models for these languages becomes a complex endeavor, hindered by limited training data and insufficient linguistic representation [9]. Low-resource languages lack the necessary data processing and ground truth data requirements for machine learning systems [10].

The urgency of mitigating offensive content in Low-Resource Indo-Aryan Languages cannot be understated. With millions of users communicating in these languages on various platforms, there is a pressing need to safeguard them from the harmful impact of offensive language. To address this need, we propose a solution based on Multilingual Toxicity Analysis, a cutting-edge approach that transcends language barriers and facilitates the detection of offensive content across multiple languages [11].

In this research, we embark on an experimental journey, aiming to harness the power of transfer learning and zero-shot transfer techniques. Our goal is to create a model capable of identifying hate speech and offensive content in Hindi, Sinhala, Gujarati, Bengali, Bodo, Assamese languages despite being primarily trained on English data through few-shot learning with relatively smaller datasets. By strategically transferring knowledge from high-resource languages like English to low-resource languages, we aim to enhance the capabilities of NLP models in identifying toxic language across linguistic boundaries.

The subsequent sections of this paper will delve into the methodology employed, including data collection and preprocessing techniques, the selection and fine-tuning of NLP models, as well as the evaluation metrics used to assess model performance. We will discuss the insights drawn from our experiments, shedding light on the model's adaptability in diverse linguistic contexts. Additionally, we will explore the implications and limitations of our approach, paving the way for further advancements in the realm of Multilingual Toxicity Analysis for low-resource languages. Ultimately, our research strives to contribute to a safer and more inclusive digital ecosystem by demonstrating that, with strategic fine-tuning, NLP models can transcend language barriers and effectively detect offensive content. By doing so, we aspire to create online spaces where offensive content is promptly identified and addressed, fostering an atmosphere of respect and tolerance among users.

## 2. Literature Review

The literature surrounding hate speech and offensive content detection has witnessed substantial growth due to the increasing concern over the proliferation of harmful online behavior [1, 3]. This literature review aims to explore existing research in the field and shed light on innovative approaches for addressing this pressing issue. By conducting a comprehensive analysis of prior studies, we seek to identify new perspectives, reveal gaps in the literature, resolve conflicting findings, and prevent duplication of effort.

Several foundational studies have contributed significantly to the development of Natural Language Processing (NLP) models as central characters in the fight against offensive content [4, 12]. DistilBERT [13] and XLM-RoBERTa [14] are two such prominent BERT [15] models that have shown promise in detecting hate speech and offensive language across diverse languages and contexts. Researchers have devoted considerable effort to honing and refining these characters, exploring their strengths and limitations in identifying toxic language patterns.

The setting of this literature review encompasses the digital landscape, where social media platforms and online communities serve as primary channels for communication. In this virtual environment, offensive content often thrives, necessitating effective content moderation strategies [16, 4]. Understanding this setting is crucial for contextualizing the challenges and opportunities for NLP models in detecting and mitigating hate speech and offensive language effectively.

The plot of this literature review revolves around the development and evaluation of NLP models tailored for hate speech and offensive content detection. Researchers have embarked on a journey to explore various embedding-classifier pairs, employing techniques such as DistilBERT with Decision Tree, Gaussian Naive Bayes, Neural Network, and XLM-RoBERTa embedder with its corresponding classifier [14, 17]. The plot thickens as researchers delve into the intricacies of these models, seeking to improve performance and generalization across diverse languages and contexts.

The overarching theme that unifies the literature is the pursuit of fostering a safer online environment and promoting linguistic inclusivity. By advancing multilingual toxicity analysis, researchers contribute to creating digital ecosystems where users from diverse linguistic backgrounds can engage in respectful and tolerant discourse [18, 2]. The theme emphasizes the significance of breaking barriers, both linguistic and cultural, using advanced NLP techniques and cross-lingual representation learning to build models that transcend language-specific limitations.

This literature review is framed within the context of a broader research landscape that spans computational linguistics, NLP, and machine learning. It situates itself at the forefront of addressing the challenges posed by offensive content in Low-Resource Indo-Aryan Languages like Sinhala and Gujarati [9, 19, 20]. By contextualizing our review within this frame, we underscore the significance of our research in bridging gaps and offering insights to the broader academic community.

The exposition in this literature review provides a comprehensive overview of the current state of research in hate speech and offensive content detection. We delve into existing studies, methodologies, and NLP models, outlining the strengths and limitations of various approaches. Through this exposition, we aim to provide a solid foundation for understanding the challenges

and potential solutions in multilingual toxicity analysis.

In conclusion, this literature review serves as a critical link between the introduction and methodology sections of our research. By exploring the foundational elements of NLP models as characters, the digital setting, the plot of NLP techniques for offensive content detection, the overarching theme of fostering a safer online environment, and the framing within the broader research landscape, we gain a deeper understanding of the landscape of multilingual toxicity analysis. Our analysis aims to identify new avenues for research, resolve conflicts in existing studies, and contribute to the collective effort of creating a safer and more inclusive digital ecosystem.

### 3. Methodology

#### 3.1. Data processing

##### 3.1.1. English

The dataset utilized was from the Kaggle jigsaw-toxic-comment-classification competition. It had the following toxicity types:

- |                 |            |                  |
|-----------------|------------|------------------|
| 1. toxic        | 3. obscene | 5. insult        |
| 2. severe_toxic | 4. threat  | 6. identity_hate |

Since they are all regarded as harmful, if a comment contained at least one of the above labels, we regarded it as toxic and combined the comments into a new super column. The dataset approximately had 150k samples of labelled data, after dropping the null values. Moreover, the data was cleaned by removing unnecessary symbols, tags, etc. For training 80% of the samples were used while 20% were stored for validation and evaluation.

##### 3.1.2. Hindi

The hindi dataset from [21] was used for hindi fine-tuning. The dataset size had 1603 training samples in text format with labels. The types of unique labels that were found after being separated from text included the following:

- |   |  |
|---|--|
| 1. __label__abusive,                                  | 9. __label__hatespeech__label__abusive,                  |
| 2. __label__abusive__label__hatespeech,               | 10. __label__hatespeech__label__abusive__label__abusive, |
| 3. __label__abusive__label__hatespeech__label__humor, | 11. __label__hatespeech__label__benign,                  |
| 4. __label__abusive__label__humor,                    | 12. __label__humor__label__abusive,                      |
| 5. __label__benign,                                   | 13. __label__humor__label__benign,                       |
| 6. __label__benign__label__abusive,                   | 14. __label__humor__label__hatespeech__label__abusive    |
| 7. __label__benign__label__humor,                     |  |
| 8. __label__hatespeech,                               |  |

We labelled the ones containing 'label\_\_abusive' or 'label\_\_hatespeech' as 'HOF' and the rest as 'NOT' where 'HOF' is encoded further as 1, and 'NOT' as 0. The train set had 1603 samples while test had 397.

### 3.1.3. Gujarati

The Gujarati dataset contains 200 tweets. Firstly [22, 23], the tweets were cleaned by removing unnecessary symbols, tags, and mentions. Moreover, the labels 'HOF' and 'NOT' in the dataset were encoded to 1 and 0 respectively. The same procedure was used for test set which contained 1196 samples.

### 3.1.4. Bengali, Assamese, Sinhala and Bodo

The datasets of Bengali [24, 23], Assamese [25, 24, 23], Sinhala [26, 23], and Bodo [24, 23] are primarily collected from Twitter, Facebook, or youtube comments, all were cleaned with the same process as Gujarati, and had the same label convention.

### 3.1.5. Findings from datasets

After analyzing each dataset, it was found that each sample had, on average, fewer than 100 tokens. Based on these findings we decided the '*max\_length*' of the model. Moreover, for tokenization '*word\_tokenize*' from 'nltk' was used to tokenize the samples, and average tokens (*avg\_tokens*) were calculated using the average formula that is:

$$avg\_tokens = \frac{\sum_{i=1}^n \sum_{w=1}^m T_w}{\sum_{k=1}^l k}$$

where:

1. n: total number of samples
2.  $T_w$ : token of a word
3. m: total words in a sample
4. l: total number of samples

## 3.2. Model and architecture

We employ the XLM-RoBERTa model, a powerful pre-trained language representation model, for the task of hate speech and offensive content detection. XLM-RoBERTa, short for Cross-lingual Language Model - RoBERTa, is an extension of the RoBERTa model that is specifically designed to handle multilingual text, and performs particularly well on resource languages[14].

The architecture of XLM-RoBERTa consists of an embeddings layer, followed by an encoder layer. The embeddings layer includes word embeddings, position embeddings, and token type embeddings, which together form the input representation for the model. The encoder layer comprises a series of XLM-RoBERTa layers, each consisting of attention mechanisms and feed-forward neural networks [14].

The attention mechanism in XLM-RoBERTa enables the model to focus on important parts of the input sequence during processing. This mechanism includes self-attention, where each

word in the input is associated with weights that determine its relevance to other words in the sequence [27].

Furthermore, XLM-RoBERTa incorporates LayerNorm and dropout layers to improve model stability and prevent overfitting. LayerNorm normalizes the hidden state of each layer, and dropout randomly deactivates certain neurons during training to prevent the model from relying too heavily on specific features [28, 29].

The classification head of XLM-RoBERTa consists of a dense layer, followed by a dropout layer, and an output projection layer. The dense layer helps in reducing the feature dimensionality, while dropout aids in regularization to avoid overfitting. The output projection layer maps the reduced features to the final output, which is a binary classification for hate speech and offensive content detection [14]. The model size is 278M parameters. However, trainable classification head has 592,130 parameters. Additionally, the maximum length of the model was set at 256 for Code-Mixed Languages and 128 for the rest based on the results of the dataset processing because the average token count was less than 100. The figure 1 below illustrates sample model pipeline. The pipeline consist of an multi-lingual embedder followed by a classifier:

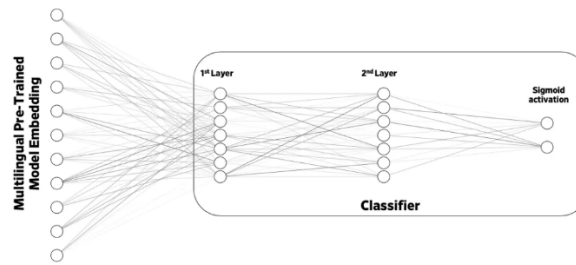


Figure 1: Model pipeline

In summary, the XLM-RoBERTa model, with its multilingual representation learning capabilities and attention-based architecture, is a robust choice for our hate speech and offensive content detection task. By leveraging its pre-trained language understanding, we aim to achieve accurate and effective results across diverse languages and contexts.

### 3.3. Model Implementation

We propose a model (the implementation can be viewed at [https://github.com/Jayveersinh-Raj/indo\\_aryan\\_clf\\_HASOC\\_2023](https://github.com/Jayveersinh-Raj/indo_aryan_clf_HASOC_2023)) which is initially trained on an English dataset and subsequently fine-tuned on different languages. By following the above approach the model can be tailored to any low resource language with relatively small dataset. The naming conventions for the models are as follows:

- **XLM-Classifier**: The radomly initialized classifier not finetuned for downstream task.
- **PolyGuard**: The English only trained model (XLM-Classifier) [30]
- **Indo-Aryan-Extension**: The PolyGuard finetuned on Hindi [31]
- **xlm-sinhala**: The Sinhala only trained model (XLM-Classifier)
- **xlm-bengali**: The Bengali only trained model (XLM-Classifier)

- **xlm-bodo**: The Bodo only trained model (XLM-Classifier)
- **xlm-assamese**: The Assamese only trained model (XLM-Classifier)

It is important to note that in all of the models above, and if fine-tuned, we utilized the **AdamW** optimizer which is an extension of the Adam optimizer using a method known as weight decay. To stop overfitting, a regularization factor called weight decay is applied to the loss function. By separating weight decay from the optimization processes, AdamW solves the weight decay problems with the original Adam optimizer, providing more reliable and efficient training. The common **hyperparameters** used are as follows:

- learning rate:  $2 \times 10^{-5}$
- epsilon:  $1 \times 10^{-8}$
- Trainable Layers: All

These hyper-parameters were chosen to balance precision in weight updates and numerical stability during the optimization process, ensuring the model's effective training and performance. Another important hyperparameter we used was epoch, which is mentioned in each of the sub-experiments in the results section. We tested different epochs on each sub-dataset to find the best performance, and we chose the epochs that performed well on the unseen test data, which helped to mitigate overfitting and underfitting.

## 4. Results

In this section, we present the results of our experiments in fine-tuning the XLM-RoBERTa model for hate speech and offensive content detection in various Indian languages. The experiments considers fixing the model while testing the performance on various datasets, and few shot capabilities of the model. The primary focus for few shot inference on languages such as Gujarati which has relatively low size leaving no room to train a raw architecture on the dataset.

We report the dataset details and performance metrics for each language below, however, it is imperative to mention that in order to avoid over-fitting we verified the outcomes by training the models for a slightly shorter number of epochs than stated below for each language.

### 4.1. English

#### Dataset Details:

- Training Dataset: Approximately 150k samples encoded as 1 for Hatespeech, and 0 for neutral
- Total Training Samples: Approximately 120k
- Test Dataset: Approximately 30k.

Model performances		
Model	epochs	F1-score
XLM-Classifier	3	0.960

## 4.2. Hindi

### Dataset Details:

- Training Dataset: Cleaned with sentences split from labels. 'Hatespeech' labeled as 'HOF' (1), the rest as 'NOT' (0).
- Total Training Samples: 1603
- Test Dataset: Similar preprocessing applied, containing 397 samples.

Model performances		
Model	epochs	F1-score
Direct inference (PolyGuard)	-	0.458
Fine-tuned (PolyGuard)	10	0.889

## 4.3. Gujarati

### Dataset Details:

- Training Dataset: Contains 200 crawled tweets in Gujarati labeled 'HOF' (1) for hate speech and 'NOT' (0) for neutral.
- Unique local slangs observed in tweets, cleaned by removing mentions and symbols.
- Test set contained 1196 tweets

Model performances			
Model	epochs	F1-score	Leaderboard Rank
Direct inference (PolyGuard)	-	0.165	-
Direct inference (Indo-Aryan-Extension)	-	0.547	-
Fine-tuned (PolyGuard)	20	0.791	-
<b>Fine-tuned (Indo-Aryan-Extension)</b>	<b>20</b>	<b>0.793</b>	<b>4</b>

## 4.4. Sinhala

### Dataset Details:

- Training Dataset: Contains 7,500 samples in Sinhala labeled 'HOF' (1) for hate speech and 'NOT' (0) for neutral.
- Test Samples: 2,500.

Model performances			
Model	epochs	F1-score	Leaderboard Rank
<b>xlm-sinhala</b>	<b>10</b>	<b>0.835</b>	<b>3</b>
Fine-tuned (Indo-Aryan-Extension)	10	0.827	-
Fine-tuned (PolyGuard)	10	0.820	-



## 4.5. Bengali

### Dataset Details:

- Training Dataset: Contains 1281 samples in Bengali labeled 'HOF' (1) for hate speech and 'NOT' (0) for neutral.
- Test samples: 320

Model performances			
Model	epochs	F1-score	Leaderboard Rank
<b>Fine-tuned(Indo-Aryan-Extension)</b>	<b>30</b>	<b>0.726</b>	<b>6</b>
<b>Fine-tuned (PolyGuard)</b>	<b>30</b>	<b>0.726</b>	<b>6</b>
xlm-bengali	30	0.634	-

## 4.6. Bodo

### Dataset Details:

- Training Dataset: Contains 1679 samples in Bodo labeled 'HOF' (1) for hate speech and 'NOT' (0) for neutral.
- Test samples: 420

Model performances			
Model	epochs	F1-score	Leaderboard Rank
Fine-tuned(Indo-Aryan-Extension)	20	0.829	-
<b>Fine-tuned (PolyGuard)</b>	<b>20</b>	<b>0.844</b>	<b>6</b>
xlm-Bodo	20	0.831	-

## 4.7. Assamese

### Dataset Details:

- Training Dataset: Contains 4036 samples in Assamese labeled 'HOF' (1) for hate speech and 'NOT' (0) for neutral.
- Test samples: 1009

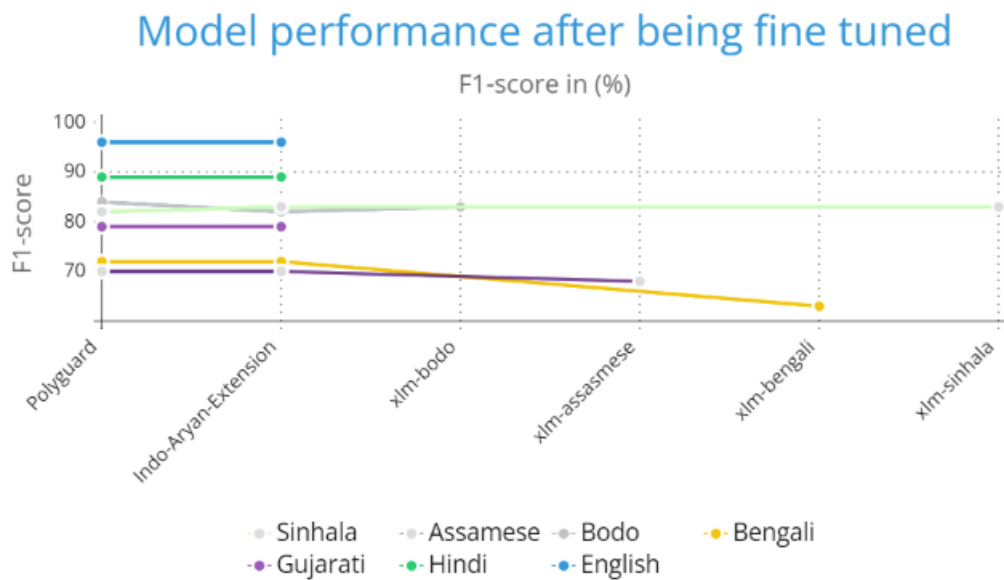
Model performances			
Model	epochs	F1-score	Leaderboard Rank
<b>Fine-tuned (Indo-Aryan-Extension)</b>	<b>30</b>	<b>0.707</b>	<b>5</b>
Fine-tuned (PolyGuard)	30	0.706	-
xlm-assamese	30	0.681	-

#### 4.8. General Analysis

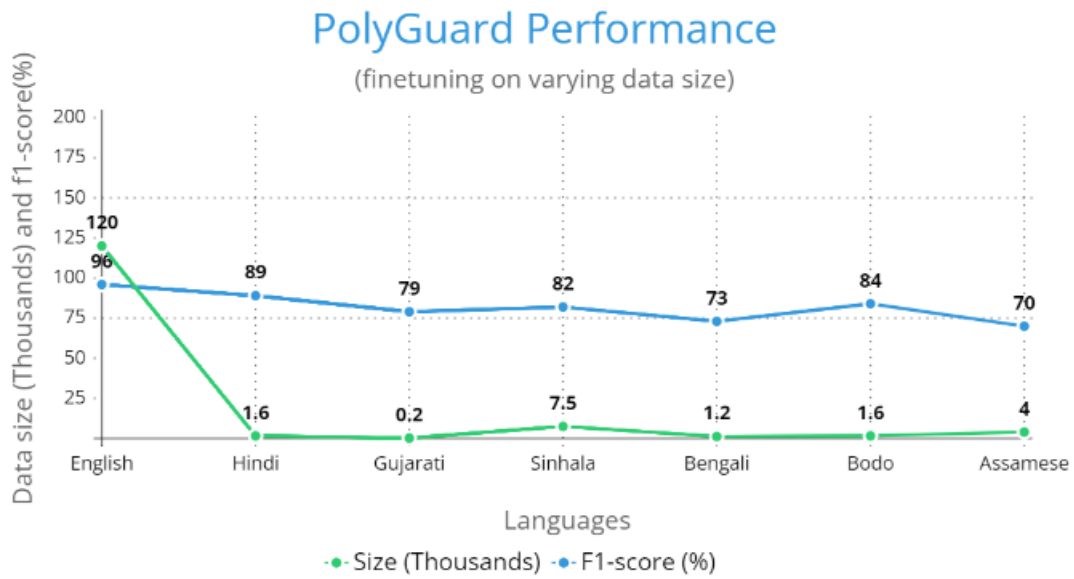
- The fine-tuning process demonstrates the model's adaptability to different languages with significantly low fine tuning samples of the target language, with F1 scores ranging from 0.70 to 0.89.
- Notably, the model performs exceptionally well in languages like Sinhala and Bodo, where F1 scores exceed 0.8.
- The performance in Assamese and Bengali datasets is moderate, suggesting room for further improvement.

These experiments demonstrate the XLM-RoBERTa model's potential for cross-lingual hate speech detection, as well as the potential of our 'PolyGuard' model, which, after being trained on 100k+ samples of English, can produce significant results after being fine-tuned with a very small dataset of a target language. The results vary across languages, emphasizing the need for language-specific adaptations and larger, diverse datasets to enhance model performance further. Future work should focus on addressing the challenges posed by low resource languages and exploring techniques to improve classification accuracy.

The following figure demonstrates that fine tuned 'PolyGuard' yields better results over multiple languages:



Moreover, the results of the fine-tuned 'PolyGuard' model on varying dataset size can be seen in the figure below:



## 5. Discussion

In this section, we delve into the insights drawn from the results obtained in our experiments. The varying performance across languages and datasets provides valuable insights into the challenges and opportunities in multilingual hate speech detection.

### 5.1. Cross-Lingual Adaptability

Our experiments have showcased the remarkable cross-lingual adaptability of the XLM-RoBERTa model. It excelled in languages like Sinhala and Bodo, where the F1 scores exceeded 0.8. This demonstrates the model's potential for effectively identifying hate speech in languages with relatively larger datasets and well-defined linguistic patterns. It aligns with prior research indicating that multilingual models can leverage their pre-trained knowledge effectively across languages [32].

### 5.2. Data Quality and Size

The performance in Gujarati is notable, with an F1 score of 0.7926. It's essential to highlight that this dataset contained local slangs and expressions that might only be discernible to native speakers. This raises questions about data quality and the importance of domain-specific knowledge in hate speech detection tasks. Additionally, the dataset sizes played a crucial role in model performance. Languages with larger datasets, such as Sinhala and Bodo, yielded higher F1 scores. Hence, expanding and diversifying datasets for low-resource languages is essential for improved model performance.

### 5.3. Room for Improvement

While our model achieved competitive results, there is room for improvement. For instance, in Bengali and Assamese datasets, where F1 scores were moderate (0.726 and 0.707 respectively), further fine-tuning, data augmentation, or specialized techniques for handling local dialects may enhance performance. Additionally, exploring more advanced model compression techniques could optimize efficiency while retaining performance [33].

### 5.4. Future Directions

Future work in cross-lingual hate speech detection should focus on:

1. Investigating techniques for better handling dialects.
2. Expanding and diversifying datasets for low-resource languages.
3. Exploring advanced fine-tuning strategies to boost model performance.
4. Addressing issues related to model efficiency and interpretability.

Our experiments underscore the adaptability and potential of the XLM-RoBERTa model in multilingual hate speech detection. However, they also shed light on the complexities and the critical role of dataset size and quality. These findings provide a foundation for further research aimed at bridging the gap in hate speech detection across diverse languages and cultures.

## 6. Conclusion

In this study, we embarked on a journey to break barriers in multilingual toxicity analysis, focusing on hate speech and offensive content detection in low-resource languages. Leveraging the formidable capabilities of the XLM-RoBERTa model, we conducted a series of experiments fine-tuning the model for various languages. The insights drawn from our endeavors provide valuable contributions to the field of cross-lingual hate speech detection.

Our results are promising, indicating the remarkable cross-lingual adaptability of the XLM-RoBERTa model. It demonstrated exceptional performance in languages such as Sinhala and Bodo, achieving F1 scores exceeding 0.8. These outcomes underscore the potential of multilingual models to effectively identify hate speech in languages with relatively larger datasets and well-defined linguistic patterns. This aligns with the vision of creating models that can transcend language barriers.

Our experiments also emphasized the critical importance of data quality and dataset size. In Gujarati, the model achieved commendable results (F1 score of 0.7926), but the dataset contained local slangs and expressions that might be understood only by native speakers. This raises questions about the significance of domain-specific knowledge in hate speech detection tasks. Additionally, languages with larger datasets, like Sinhala and Bodo, yielded higher F1 scores, highlighting the importance of dataset expansion and diversification for low-resource languages.

Despite these challenges, our model has exhibited its potential to deliver competitive performance in multilingual hate speech detection. With access to balanced datasets and further fine-tuning efforts, we believe our model can surpass competitors in the field, making significant strides in mitigating hate speech and offensive content in diverse linguistic landscapes. This

study has also demonstrated the generalizability of our model to different datasets, paving the way for its broader application.

In conclusion, our journey in breaking barriers through multilingual toxicity analysis has yielded promising results. The XLM-RoBERTa model's adaptability, coupled with the insights gained from our experiments, sets the stage for future research and advancements in hate speech detection across languages and cultures. As we continue to refine and expand our model, we are one step closer to fostering more inclusive and respectful online spaces for all.

## References

- [1] X. Zhang, F. Luo, X. Hu, Detecting hate speech on twitter using a convolution-gru based deep neural network, *Proceedings of the 27th international conference on computational linguistics* (2018) 2561–2571.
- [2] T. Davidson, D. Warmley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, *Proceedings of the 11th international conference on web and social media* (2017) 512–515.
- [3] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, *NAACL-HLT* (2016) 88–93.
- [4] A. Schmidt, M. Wiegand, A survey of hate speech detection using natural language processing, *Proceedings of the fifth international workshop on natural language processing for social media* (2017) 1–10.
- [5] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Eighth international conference on weblogs and social media* (2014) 216–225.
- [6] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford 1* (2009) 2009.
- [7] Z. Waseem, D. Hovy, Understanding abuse: A typology of abusive language detection subtasks, *Proceedings of the first workshop on abusive language online* (2017) 1–10.
- [8] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, De La Salle University, Manila, Philippines, 2022*, pp. 853–865. URL: <https://aclanthology.org/2022.paclic-1.94>.
- [9] M. Geva, Y. Goldberg, Contextualizing hate speech classifiers with post-hoc explanations, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019) 2736–2741.
- [10] D. Nkemelu, H. Shah, I. Essa, M. L. Best, Tackling hate speech in low-resource languages with context experts, 2023. [arXiv:2303.16828](https://arxiv.org/abs/2303.16828).
- [11] T. Schuster, J. Bastings, Y. Belinkov, S. Goldwater, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019*, pp. 3500–3508.

- [12] W. Liu, S. Singh, M. Gardner, P. Kohli, Deep learning for extreme multi-label text classification, arXiv preprint arXiv:1905.07342 (2019).
- [13] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.
- [14] A. Conneau, G. Lample, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 8440–8451.
- [15] K. Ghosh, A. Senapati, U. Garain, Baseline bert models for conversational hate speech detection in code-mixed tweets utilizing data augmentation and offensive language identification in marathi, in: Fire, 2022. URL: <https://api.semanticscholar.org/CorpusID:259123570>.
- [16] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2013) 1250–1259.
- [17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
- [18] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, 2021. arXiv:2102.08886.
- [19] A. Vadesara, P. Tanna, Corpus building for hate speech detection of gujarati language, in: International Conference on Soft Computing and its Engineering Applications, Springer, 2022, pp. 382–395.
- [20] H. Sandaruwan, S. Lorensuhewa, M. Kalyani, 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, IEEE, 2019, pp. 1–8.
- [21] V. K. Jha, H. P. V. P. N, V. Vijayan, P. P, Dhot-repository and classification of offensive tweets in the hindi language, Procedia Computer Science 171 (2020) 2324–2333. doi:<https://doi.org/10.1016/j.procs.2020.04.252>.
- [22] 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.
- [23] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.
- [24] K. Ghosh, A. Senapati, A. S. Pal, Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
- [25] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi, A. Senapati, Transformer-based hate speech detection in assamese, in: 2023 IEEE Guwahati Subsection Conference (GCON), 2023, pp. 1–5. doi:10.1109/GCON58516.2023.10183497.
- [26] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, 2022. arXiv:2212.00851.

- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* 30 (2017) 5998–6008.
- [28] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, *arXiv preprint arXiv:1607.06450* (2016).
- [29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research* 15 (2014) 1929–1958.
- [30] J. Raj, Polyguard model, 2023. URL: <https://huggingface.co/Jayveersinh-Raj/PolyGuard>.
- [31] J. Raj, Indo-aryan-abuse-detection, 2023. URL: <https://huggingface.co/Jayveersinh-Raj/Indo-Aryan-abuse-detection>.
- [32] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, 2020. *arXiv:2008.00401*.
- [33] A. Berthelot, T. Chateau, S. Duffner, C. Garcia, C. Blanc, Deep model compression and architecture optimization for embedded systems: A survey, *Journal of Signal Processing Systems* 93 (2021) 863–878.