

Hate and Offensive Content Identification in Indo-Aryan Languages using Transformer-based Models

Olumide Ebenezer Ojo^{1,2,†}, Olaronke Oluwayemisi Adebajji¹, Hiram Calvo¹, Alexander Gelbukh¹, Anna Feldman² and Grigori Sidorov¹

¹*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*

²*Montclair State University, USA*

Abstract

Open exchange of hate speech, insults, derogatory remarks, and obscenities on social media platforms can undermine objective discourse and facilitate radicalization by spreading propaganda and exposing people to danger. People who have been targeted by these offensive and hateful content often experience physiological effects as a result. In this work, we present our models for detecting hate speech and offensive content in two Indo-Aryan languages submitted to HASOC 2023. Although Gujarati and Sinhala are considered low-resource languages, our models demonstrated commendable accuracy in detecting hate speech after fine-tuning them with language-specific hate speech datasets. Our experiments employed and fine-tuned two transformer models, namely DistilBERT and mBERT, and we show that these transformer models were effective in detecting hate speech in Indo-Aryan texts. mBERT achieved the macro F1-score of 0.6 in the Sinhala text and excelled further with a score of 0.8 in the Gujarati text classification.

Keywords

Hate Speech, Offensive Content, Gujarati, Sinhala, Transformers

1. Introduction

With unparalleled global connectivity and communication, the emergence of hate speech and offensive content on social media platforms and other online spaces has become an alarming concern. While this technological progress has brought numerous benefits, it has also created significant challenges in the form of hate speech and offensive content in online spaces [1, 2, 3, 4, 5, 6]. The widespread dissemination of harmful language, discriminatory rhetoric, and offensive materials has not only tainted online discourse, but has also raised serious social concerns. Addressing this issue is imperative to ensure the safety, inclusivity, and well-being of users and communities that participate in different online platforms.


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

✉ olumideoea@gmail.com (O. E. Ojo); olaronke.oluwayemisi@gmail.com (O. O. Adebajji); hcalvo@cic.ipn.mx (H. Calvo); gelbukh@cic.ipn.mx (A. Gelbukh); feldmana@montclair.edu (A. Feldman); sidorov@cic.ipn.mx (G. Sidorov)

🆔 0000-0003-3500-5218 (O. E. Ojo); 0000-0002-7412-6277 (O. O. Adebajji); 0000-0003-2836-2102 (H. Calvo); 0000-0001-7845-9039 (A. Gelbukh); 0000-0003-4146-5990 (A. Feldman); 0000-0003-3901-3522 (G. Sidorov)



© 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)

A significant threat to social media users is hate speech that denigrates, targets, or promotes violence against people or groups. Its impact extends beyond the virtual world, often spilling into the real world with real consequences. Managing social media platforms has become increasingly difficult due to offensive content, which encompasses a range of harmful language and behaviors. Efforts to combat hate speech and offensive content have been ongoing and research into effective detection and mitigation methods has gained considerable traction. The use of NLP and machine learning has led to the development of automated solutions that can identify hate speech and offensive content quickly and accurately.

The linguistic milieu of the Indo-Aryan region [7, 8], which includes languages spoken in South Asia, poses a distinctive challenge when it comes to identifying hate speech. Due to the diversity of dialects in Indo-Aryan languages, comprehensive hate speech detection tools have been difficult to develop due to the lack of linguistic resources and annotated datasets. This article aims to contribute to ongoing efforts to combat hate speech and offensive content in these Indo-Aryan languages. We explore the application of BERT-based approaches, specifically mBERT and DistilBERT, to enhance hate speech detection in these languages. By fine-tuning these models on language-specific hate speech datasets, we aim to provide effective solutions that can foster healthier online conversations.

The HASOC (Hate Speech and Offensive Content) competition was created to promote research in automatically identifying hate speech and offensive content across diverse languages. As our society becomes increasingly reliant on technology, this competition aims to promote the development of tools and techniques that can help combat online hate speech. The primary goal of the HASOC competition is to motivate researchers to design and develop automated systems that can detect hate speech and offensive content accurately. One of the distinctive features of HASOC is its focus on multiple languages, thereby promoting research in low-resourced languages. Although, much hate speech detection efforts have traditionally focused on English, HASOC acknowledges that hate speech is a global problem.

In the fifth edition of the HASOC competition, the organizers provided labeled datasets in Sinhala [9], an Indo-Aryan language spoken in Sri Lanka, and Gujarati, another Indo-Aryan language spoken by around 50 million people in India. Task 1 of the competition focuses on the use of NLP techniques to detect hate speech and offensive language in various languages. These datasets consist of hate speech, offensive language, and non-offensive content, labeled Hate and Offensive (HOF) or Non-Hate and Offensive (NOT). HASOC typically evaluates participating models based on standard metrics, including macro F1, macro precision, and macro recall scores.

Various tasks related to text classification [10, 11, 12, 13, 14, 15, 16], including those focused on detecting hate speech and offensive content [4, 3, 5, 1, 2] rely on NLP techniques, highlighting the need for nuanced models to address these tasks effectively. In this paper, we explore transformer-based models [17, 18] for these classification tasks, which have consistently exceeded existing baselines and established themselves as state-of-the-art solutions. In the following sections, we discuss related work in hate speech detection, detail our methodology, present experimental results, and discuss the implications of our findings. This research underscores the importance of using state-of-the-art NLP techniques to address the pressing challenge of hate speech and offensive content in this day and age, particularly in low-resource settings.

2. Related Works

Different machine learning models, including transformer-based models, have been used to address the offensive content and hate speech detection task. In [19], the authors presented an innovative transformer-based framework capable of managing various tasks, including the recognition of aggression and hate, misogynistic aggression, the identification of offensive hate content, and the detection of emotions. This proposed approach exceeded established benchmarks in multiple languages, including emotion detection to improve performance. Furthermore, the article outlines potential avenues for future research, including the use of task-specific lexicons, the incorporation of external knowledge sources, the examination of the influence of sexual and gender identities on system efficacy, and the exploration of various task loss weightings for optimal performance.

The emerging field of text-hatred speech detection, particularly in the context of the Assamese language, was studied in [20]. The rapid expansion of social media has highlighted the urgency of recognizing and dealing with offensive content that can quickly spread and possibly provoke violence. Detecting hate speech in the multilingual Indian context is challenging, and two significant contributions were made. First, they created a labeled dataset of 4,000 Assamese sentences for hate speech detection. Second, they fine-tuned existing models (mBERT cased and Bangla-BERT) using this Assamese dataset to effectively detect hate speech. Their work represents a pioneering effort in the detection of hate speech in the Assamese language, addressing a critical need in the field of NLP.

The proliferation of social networks has led to a rise in verbal abuse and hatred, particularly on platforms like Twitter. In [21], a dataset of 38K Persian hate and offensive tweets was created using keyword-based selection strategies to detect offensive language in Persian text. Lexicons for offensive language and targeted hate groups were gathered through crowd-sourcing, and the data set was manually annotated by multiple annotators. The authors also examined the bias of the dataset and mitigated its impact on the performance of the language model, achieving a significant reduction in bias with minimal loss in the F1 score. The study applied various machine learning methods and Transformer-based models to the dataset, with Transformer models proving more efficient in detecting offensive content. The study paves the way for comprehensive research on the offensive language of the Persian language and its various aspects.

In their study, [22] investigated the efficacy of transformer-based language models, including BERT, RoBERTa, ALBERT, and DistilBERT, in the task of detecting hate speech on established Indian datasets like HASOC-Hindi (2019), HASOC-Marathi (2021), and Bengali hate speech (BenHateSpeech). Traditional deep learning methods struggle to detect hate speech when hate-related terms are concealed within a sophisticated language. Transformer-based multilingual models such as MuRILBERT and XLM-RoBERTa were compared with monolingual models such as NeuralSpaceBERT-Hi (Hindi), MahaBERT (Marathi), and BanglaBERT (Bengali). The results indicated that MahaBERT excels on HASOC-Marathi, while MuRILBERT performs best on HASOC-Hindi and BenHateSpeech. Their study also conducts cross-language evaluations and highlights the scarcity of research in Indian languages such as Hindi, Bengali, Marathi, Tamil, and Malayalam. The authors successfully explored various transformer-based models in Indian languages, comparing monolingual and multilingual models for hate speech detection, and

underscore the importance of context in multilingual models, with different models excelling in different Indian language datasets.

The challenges faced by social media platforms in moderating content quickly lead to the abuse of these platforms. Cyberbullying, which occurs on online platforms, has real-world consequences such as depression and suicide attempts. [23] conducted a comprehensive survey of more than 70 studies on automatic detection of cyberbullying in low-resource languages, identifying research gaps, including the lack of clear definitions, biases in data acquisition, and annotation problems. The authors propose suggestions for improving research in this area, published a dataset on cyberbullying in the Chittagonian dialect of Bangla, and offer machine learning solutions. The analysis revealed the limitations of cyberbullying detection in low-resource language research, particularly in dataset quality and data imbalance.

A novel multilingual hate speech analysis dataset, called LAHM, was created by [24]. The dataset addressed various types of hate speech in five domains: Abuse, Racism, Sexism, Religious Hate, and Extremism. This is a pioneering effort, as it is the first dataset to tackle the identification of hate speech in these domains and languages simultaneously. The paper explains how the dataset was created, annotated at different levels, and used to test state-of-the-art multilingual and multitask learning approaches. It evaluates the dataset in various classification settings, including monolingual, cross-lingual, and machine translation classification, comparing it with existing English datasets. The authors discuss the potential for creating large-scale hate speech datasets using this approach and improving hate speech detection in general. LAHM is described as one of the largest datasets of its kind, containing nearly 300k tweets in six languages and five domains. It facilitates cross-lingual abusive language detection and allows for the exploration of language and domain shifts.

In this section, we provide an overview of the prevailing trends in hate speech detection research, including multilingual approaches, deep learning methods, and fine-tuning techniques. Our study contributes to the detection of hate speech through the development of transformer-based models for identifying hate speech and offensive content in Sinhala and Gujarati text.

3. Dataset Description

Originating from the Indian state of Gujarat, the Gujarati language has a rich literary history. The Gujarati dataset consists of tweets that comprise hate/offensive content and non-hate content. These tweets provide a snapshot of contemporary issues, sentiments, and potential biases present within the Gujarati-speaking online community. Sinhala, the native language of the Sinhalese people, is the major language of Sri Lanka. The Sinhala dataset, like Gujarati, captures the nuances of hate speech and offensive content in the digital space. By analyzing this dataset, we gained insights into the socio-political dynamics and potential sources of contention within the Sinhala-speaking community. Using advanced transformer-based models, we address this specific challenge in Task 1 of the HASOC 2023 competition [25]. Subtasks A and B focused on detecting hate speech and offensive language in Gujarati and Sinhala languages. These datasets will be used to develop machine learning models capable of detecting and mitigating hate speech in regional languages, thus promoting positive online interactions and reducing harm.

3.1. Task 1A: Identifying hate, offensive and profane content in Sinhala text

In Task 1A, the task was to identify hate and offensive content in Sinhala, an Indo-Aryan language with limited linguistic resources. A key element of this task is the classification of tweets specific to the Sinhala language into two distinct categories: Hate and Offensive (HOF) and Non-Hate and Offensive (NOT). The dataset used for this task is derived from the Sinhala Offensive Language Detection dataset [9]. As an official language in Sri Lanka, Sinhala is spoken by more than 17 million people, and in this unique linguistic setting, HASOC introduces its inaugural shared task for the processing of the Sinhala language. This task adopts a coarse-grained binary classification approach, with participating systems being required to categorize tweets into either a Hate and Offensive (HOF) or a Non-Hate and Offensive (NOT). A tweet that falls into the NOT category does not contain hate speech or offensive content, while a tweet that falls into the HOF category does include elements of hate and offensiveness.

3.2. Task 1B: Identifying hate, offensive and profane content in Gujarati text

The objective of Task 1B is to detect hate speech and offensive material in Gujarati, another Indo-Aryan language that is low-resource in nature and spoken by a population of approximately 50 million people throughout the country. As with the Sinhala version, participants are asked to categorize tweets into two categories: Hate and Offensive (HOF) and Non-Hate and Offensive (NOT). Gujarati is one of the 22 official languages in India, and HASOC 2023 extends its reach to encompass the multifaceted challenges of detecting hate speech and offensive content there. Participants are tasked with accurately categorizing tweets into two mutually exclusive categories: Hate and Offensive (HOF) and Non-Hate and Offensive (NOT). Accordingly, NOT represent tweets that contain no hate speech or offensive content, while HOF indicates tweets that contain elements of offensiveness and hate.

The statistics of the dataset are shown in Table 1 below.

| Dataset | Label | Total |
|----------------------------------|-----------------------------|-------|
| Gujarati Dataset - Training Data | HOF (Hate or Offensive) | 100 |
| | NOT (Not Hate or Offensive) | 100 |
| Gujarati Dataset - Test Data | Total | 1,196 |
| Sinhala Dataset - Training Data | HOF (Hate or Offensive) | 3,176 |
| | NOT (Not Hate or Offensive) | 4,324 |
| Sinhala Dataset - Test Data | Total | 2,500 |

Table 1: Dataset Statistics

4. System Description

Our system uses deep learning techniques and leverages pre-trained language models to perform this classification task.

4.1. Task 1A - Sinhala Text Classification

DistilBERT [26], a distilled version of BERT, and the multilingual variant of the bidirectional encoder representations of transformers [27] models were used. Using input tokens and attention masks, these models generated label predictions through a fully connected layer that emphasized relevant information. We pre-processed and tokenized the text data and constructed data loaders to handle the batch of input data efficiently for both training and testing, while ensuring padding and truncation to a maximum length of 512 tokens. We adjusted the models' parameters and applied a technique called backpropagation [28]. We were able to efficiently handle the additional computations needed during reversible training, while also calculating gradients, in order to handle the additional workload required for both activations and gradients. This helps the models learn and adjust its internal parameters to better fit the data, ultimately improving its ability to make accurate classifications. We employed the Adam optimizer with a learning rate of $3e-5$ and introduced a learning rate scheduler to dynamically adjust the learning rate during training. The models were trained for a specified 12 epochs.

Table 2 gives a summary of the hyperparameters used to train the models for the classification of the Sinhala text.

| Hyperparameter | mBERT | DistilBERT |
|---------------------------|------------------------------|------------------------------------|
| Tokenizer | bert-base-multilingual-cased | distilbert-base-multilingual-cased |
| Number of Training Epochs | 12 | 12 |
| Batch Size | 16 | 16 |
| Warm-up Steps | 500 | 500 |
| Weight Decay | 0.01 | 0.01 |
| Learning Rate | $3e-5$ | $3e-5$ |

Table 2: Hyperparameters Used for Model Training

4.2. Task 1B - Gujarati Text Classification

To address the issue of having a small dataset for training, we fine-tuned by integrating the calculated class weights into the cross-entropy loss function. The objective here was to make the few classes represented more significant during training by assigning greater importance to them in the loss calculation. By doing this, the model focused its efforts on learning these few classes. We applied mBERT for the few-shot classification task [29]. This model employs a pre-trained transformer encoder, initially trained with two primary objectives: masked token prediction and next sentence prediction. In the context of this binary classification task, both the text and the associated labels were embedded into the input. This involved the training of the model and its subsequent fine-tuning to align with our target objective.

Furthermore, we also trained and evaluated the dataset using the DistilBERT model for the few-shot classification task. With the model's tokenizer, we truncated/padded the text to 128

tokens and encoded it numerically. The labels were encoded, and class weights were assigned to address the few data classes. We initialized the model with weighted cross-entropy loss and our training parameters include 12 epochs, batch sizes (16 for training), warm-up steps (500), weight decay (0.01), and learning rate ($3e-5$). We trained the model, incorporating early stopping and were able to use the best model to predict the labels, with inverse transformation.

5. Experimental Results

We present the experimental results of the identification of hate and offensive content in Sinhala and Gujarati text using transformer-based models. Our experiments were carried out in a manner that followed the methodology outlined in the previous section. We analyze the performance metrics of two pre-trained language models, DistilBERT and mBERT, for classifying text in two distinct languages, Sinhala and Gujarati. The evaluation is based on macro precision, macro recall, and macro F1-score. For the classification of Sinhala text, both the DistilBERT and mBERT models demonstrate similar performance across all three metrics. These results indicate that both models exhibit balanced performance in correctly identifying classes within the Sinhala text data. The similarity in performance suggests that, for Sinhala text classification, DistilBERT and mBERT may be considered comparable choices. In the context of the few-shot classification task for Gujarati text, there are notable distinctions in the performance of the two models under consideration. DistilBERT consistently maintains macro precision, macro recall, and macro F1-scores at approximately 0.51. On the contrary, mBERT demonstrates a significantly improved performance when tasked with classifying Gujarati text. It attains a macro precision of 0.77, macro recall of 0.74, and a macro F1-score of 0.75. These results suggest that mBERT excels at accurately identifying classes within Gujarati text data, demonstrating its superior performance compared to DistilBERT in this specific language classification task. Table 3 presents an overview of the performance metrics achieved by DistilBERT and mBERT on the Sinhala and Gujarati test datasets. Evaluation metrics, including macro precision, macro recall, and macro F1 score, show the models’ effectiveness in handling hate speech and offensive content detection tasks across these languages”.

| Sinhala Text | | | |
|----------------------|-----------------|--------------|----------|
| Model | Macro Precision | Macro Recall | Macro F1 |
| DistilBERT | 0.51 | 0.51 | 0.51 |
| mBERT | 0.56 | 0.56 | 0.56 |
| Gujarati Text | | | |
| Model | Macro Precision | Macro Recall | Macro F1 |
| DistilBERT | 0.51 | 0.51 | 0.51 |
| mBERT | 0.77 | 0.74 | 0.75 |

Table 3: Model Performance Metrics for Hate and Offensive Detection

6. Conclusion

In light of our experimental findings, it is evident that the capability of transformer-based models varies with the language of the text data. For Sinhala text classification, both DistilBERT and mBERT demonstrate comparable performance, making either model a viable choice for tasks within this language. However, when focusing on Gujarati text, clear performance differences emerge. While DistilBERT's metrics hover around 0.51 across all evaluated areas, mBERT displays robust performance, with values exceeding 0.74 in the metrics considered. In order to tackle classification tasks in Gujarati text, mBERT emerges as the more effective tool. Detail metrics, as shown in Table 3, further underscore the importance of language-specific model evaluations, ensuring optimal results in hate speech and offensive content detection efforts. The erratic performance of the DistilBERT model can be attributed to its lack of training in Indo-Aryan languages. In our future research, we intend to improve our analysis by using models trained on these languages and to expand our study to more low-resource languages, in order to gain a better understanding of transformer models' versatility across a wide range of linguistic contexts.

Acknowledgments

This work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20230140, 20232080 and 20231567 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

- [1] C. Sinyangwe, D. Kunda, W. P. Abwino, Detecting hate speech and offensive language using machine learning in published online content, *Zambia ICT Journal* 7 (2023) 79–84.
- [2] S. Shubhang, S. Kumar, U. Jindal, A. Kumar, N. R. Roy, Identification of hate speech and offensive content using bi-gru-lstm-cnn model, in: *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, IEEE, 2023, pp. 536–541.
- [3] I. Priyadarshini, S. Sahu, R. Kumar, A transfer learning approach for detecting offensive and hate speech on social media platforms, *Multimedia Tools and Applications* (2023) 1–27.
- [4] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, N. Crespi, Hate speech and offensive language detection using an emotion-aware shared encoder, *arXiv preprint arXiv:2302.08777* (2023).
- [5] O. E. Ojo, T. H. Ta, A. Gelbukh, H. Calvo, G. Sidorov, O. O. Adebajji, Automatic hate speech detection using deep neural networks and word embedding, *Computacion y Sistemas* 26 (2022) 1007–1013.

- [6] J. Armenta-Segura, C. J. Núñez-Prado, G. O. Sidorov, A. Gelbukh, R. F. Román-Godínez, Ometeotl@multimodal hate speech event detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained BERT models over text, in: A. Hürriyetoğlu, H. Tanev, V. Zavarella, R. Yeniterzi, E. Yörük, M. Slavcheva (Eds.), Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 53–59. URL: <https://aclanthology.org/2023.case-1.7>.
- [7] K. Talukdar, S. K. Sarma, Parts of speech taggers for indo aryan languages: A critical review of approaches and performances, in: 2023 4th International Conference on Computing and Communication Systems (I3CS), IEEE, 2023, pp. 1–6.
- [8] A. Arora, A. Farris, S. Basu, S. Kolichala, Jambu: A historical linguistic database for south asian languages, arXiv preprint arXiv:2306.02514 (2023).
- [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] O. O. Adebajji, I. Gelbukh, H. Calvo, O. E. Ojo, Sequential models for sentiment analysis: A comparative study, in: Advances in Computational Intelligence-21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Proceedings, Springer Science and Business Media Deutschland GmbH, 2022, pp. 227–235.
- [11] M. Tash, J. Armenta-Segura, Z. Ahani, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma@dravidianlangtech: Convolutional neural networks for studying correlation between lexical features and sentiment polarity in tamil and tulu languages, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 180–185.
- [12] H. T. Ta, O. E. Ojo, O. O. Adebajji, H. Calvo, A. F. Gelbukh, The combination of bert and data oversampling for answer type prediction., in: SMART@ ISWC, 2021, pp. 1–13.
- [13] M. Shahiki-Tash, J. Armenta-Segura, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org, 2023.
- [14] O. E. Ojo, A. Gelbukh, H. Calvo, O. O. Adebajji, G. Sidorov, Sentiment detection in economics texts, in: Mexican International Conference on Artificial Intelligence, Springer, 2020, pp. 271–281.
- [15] O. Ojo, A. Gelbukh, H. Calvo, O. Adebajji, Performance study of n-grams in the analysis of sentiments, Journal of the Nigerian Society of Physical Sciences (2021) 477–483.
- [16] O. E. Ojo, O. O. Adebajji, A. Gelbukh, H. Calvo, A. Feldman, Medai dialog corpus (medic): Zero-shot classification of doctor and ai responses in health consultations, 2023. arXiv: 2310.12489.
- [17] O. E. Ojo, O. O. Adebajji, H. Calvo, D. O. Dieke, O. E. Ojo, S. E. Akinsanya, T. O. Abiola, A. Feldman, Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model, 2023. arXiv: 2310.09661.
- [18] O. E. Ojo, H. T. Ta, A. Gelbukh, H. Calvo, O. O. Adebajji, G. Sidorov, Transformer-based approaches to sentiment detection, in: Recent Developments and the New Directions of

Research, Foundations, and Applications: Selected Papers of the 8th World Conference on Soft Computing, February 03–05, 2022, Baku, Azerbaijan, Vol. II, Springer, 2023, pp. 101–110.

- [19] S. Ghosh, A. Priyankar, A. Ekbal, P. Bhattacharyya, A transformer-based multi-task framework for joint detection of aggression and hate on social media data, *Natural Language Engineering* (2023) 1–21.
- [20] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi, A. Senapati, Transformer-based hate speech detection in assamese, in: *2023 IEEE Guwahati Subsection Conference (GCON)*, IEEE, 2023, pp. 1–5.
- [21] E. Kebriaei, A. Homayouni, R. Faraji, A. Razavi, A. Shakery, H. Faili, Y. Yaghoobzadeh, Persian offensive language detection, *Machine Learning* (2023) 1–21.
- [22] K. Ghosh, 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*, 2022, pp. 853–865.
- [23] T. Mahmud, M. Ptaszynski, J. Eronen, F. Masui, Cyberbullying detection for low-resource languages and dialects: Review of the state of the art, *Information Processing & Management* 60 (2023) 103454.
- [24] A. Yadav, S. Chandel, S. Chatufale, A. Bandhakavi, Lahm : Large annotated dataset for multi-domain and multilingual hate speech identification, 2023. [arXiv:2304.00913](https://arxiv.org/abs/2304.00913).
- [25] 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.
- [26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. [arXiv:1910.01108](https://arxiv.org/abs/1910.01108).
- [27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [28] C. Zhou, H. Zhang, Z. Zhou, L. Yu, Z. Ma, H. Zhou, X. Fan, Y. Tian, Enhancing the performance of transformer-based spiking neural networks by improved downsampling with precise gradient backpropagation, *arXiv preprint arXiv:2305.05954* (2023).
- [29] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. A. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, *Advances in Neural Information Processing Systems* 35 (2022) 1950–1965.