

Examining Hate Speech Detection Across Multiple Indo-Aryan Languages in Tasks 1 & 4

Gyandeep Kalita^{1,†}, Eisha Halder^{1,†}, Chetna Taparia^{1,†}, Advaita Vetagiri^{1,*} and Dr. Partha Pakray¹

¹National Institute of Technology Silchar

Abstract

Hate speech continues to be a pressing concern in online social media (OSM) platforms, necessitating effective automated detection systems. In this paper, we propose a unified approach, encompassing both Task 1 & 4, to tackle the challenge of hate speech recognition within the HASOC 2023 framework. It addresses the complexities of multilingual OSM by employing cutting-edge Natural Language Processing (NLP) techniques and leveraging powerful language models put forward by team *CNLP-NITS-PP*. The key objective is optimising precision-recall trade-offs in hate speech detection, spanning English and Indo-Aryan languages. The empirical results demonstrate the effectiveness of our approach in isolating explicit signs of hate speech, emphasizing model efficiency, interpretability, and the importance of diverse linguistic nuances in creating safer online environments. This integrated work sets the stage for advancements in hate-span detection and underlines the significance of fostering responsible and inclusive online conversations across various language environments.

Keywords

Online social media, Multilingual, Natural Language Processing, CNN, BiLSTM, BERT, GPT-2, Named Entity Recognition.

1. Introduction

Social media platforms such as Twitter and Facebook have become integral to modern life, providing a global platform for individuals to express themselves. However, the openness of these platforms has also led to the proliferation of harmful content, including hate speech and harassment [1]. This has underscored the need for automated systems to identify and address abusive language in online conversations [2] [3].

Detecting offensive content is challenging due to its diverse linguistic forms, necessitating context-aware models to pinpoint hateful or abusive text snippets [4]. Additionally, implicit forms of hate speech require the deduction of pragmatic implications [5].

The spread of hate speech and inflammatory language on social media platforms is a major worldwide problem in today's digital age, as communication plays a crucial role in determining

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

*Corresponding author.


†These authors contributed equally for Task 1 & 4.

✉ gyandeepkalita1@gmail.com (G. Kalita); eishashalder@gmail.com (E. Halder); chetna.taparia@gmail.com (C. Taparia); advaita21\protect1_rs@cse.nits.ac.in (A. Vetagiri); partha@cse.nits.ac.in (Dr. P. Pakray)

🌐 <https://parthapakray.com/> (Dr. P. Pakray)

🆔 0000-0002-0651-4171 (A. Vetagiri); 0000-0003-3834-5154 (Dr. P. Pakray)

© 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)

public debate. Low-resource languages like Sinhala, Gujarati, Bengali, Bodo, and Assamese, which have received little attention in the field of Natural Language Processing (NLP), are severely affected by this problem.

Our research activities cover a range of tasks for identifying harmful and hateful content in these underrepresented languages¹. In Task 1, we tackle Sinhala (Task 1A), a language with a unique alphabet and intricate grammatical structures, and further broaden our emphasis to Gujarati (Task 1B), where a dearth of labelled data poses a significant obstacle. Task 4 extends the study's horizons by including Bengali, Bodo, and Assamese [6] [7]. These languages, which are rich in cultural richness and legacy, have generally been disregarded in NLP research, especially when it comes to the identification of hate speech. Our study uses statistics painstakingly gathered from social media sites to use binary classification to characterize material as hate/offensive or not.

The importance of our work lies in its role in safeguarding cultural identities and developing secure online environments for these language speakers. To respond to the complexities of hate speech in different linguistic and cultural contexts, we use cutting-edge NLP approaches, language-specific feature engineering, and pre-processing. Additionally, we investigate how models developed for languages with abundant resources may be applied to languages with limited resources to improve hate speech identification.

Through this extensive study project, we hope to advance responsible digital communication, better understand how to identify hate speech in different linguistic contexts, and create a more welcoming online space for every language.

2. Literature Review

Hate speech detection in fairly low-resourced languages such as Sinhala and regional Indian languages has recently attracted research attention [8]. With the proliferation of user-generated content on social media platforms, there is an urgent need to identify and moderate hateful and offensive content in these regional languages (Mathew et al., 2021)[9].

For the Sinhala Language, a few research works have been conducted. (Munasinghe et al., 2022)[10] contributed an annotated dataset of Sinhala Tweets annotated into Hate or Non-Hate. They also developed and compared the performance of different architectures such as CNN, LSTM, BiGRU and an ensemble of various other Deep Learning architectures. (Sandaruwan et al., 2019)[11] also contributed a labeled dataset containing texts from Facebook and YouTube for Hate Detection in Sinhala and compared classification results using simple Machine Learning Classifiers such as SVMs, MNB, RFDT, etc. The SOLD: Sinhala Offensive Language Dataset, a labeled dataset for Offensive content detection in Sinhala, was contributed by (Ranasinghe et al., 2022)[12], which also forms the basis for the dataset provided for the Task 1A of HASOC 2023.

In case of Indian Languages, prior work on hate speech detection has concentrated primarily on Hindi and Malayalam. For instance, (Bohra et al., 2018)[13] presented a dataset for hate speech identification in Hindi-English code-mixed social media text. They tested various classification models, including fastText, CNN, GRU, and LSTM

¹Github Repository

For Gujarati, (Khurana et al., 2022)[14] contributed a novel model to detect hate comments in 13 Indian languages that included Gujarati based on XLM-RoBERTa (XLM-R) using the Moj Multilingual Abusive Comment Identification dataset.

For the Bengali language, (Das et al., 2020)[15] compiled a labelled dataset of YouTube comments for Bengali hate speech recognition. They compared machine learning models like SVM, NB and deep learning architectures like CNN, GRU, and capsule networks.

Assamese is a relatively low-resource language. (Ghosh et al., 2023)[16] contributed a dataset for binary hate classification in Assamese and described an approach for hate detection using various BERT models.

For the Bodo language, there is limited prior research. The HASOC 2023 shared task provides the pioneering benchmark hate speech detection dataset in Bodo. This will encourage further research in this low-resource language (Chakravarthi et al., 2021)[17]

(Vetagiri et al., 2023a)[18] leveraged GPT-2 to automatically classify online sexist content. Their work demonstrates the potential of sizable pre-trained language models for hate speech detection. In another work, (Vetagiri et al., 2023b)[4] proposed an approach using CNN-BiLSTM and domain-specific embeddings for online sexism prediction.

Much previous work has relied on machine learning and deep neural networks. But these necessitate substantial labelled datasets, which are scarce for low-resource languages. Recent emphasis has focused on multilingual models such as mBERT, which can leverage data from high-resource languages. Domain adaptation approaches have also proven effective in adapting models trained on English data.

The HASOC 2023 shared tasks furnishes standard labeled benchmark datasets for hate speech detection. This will catalyze research in these languages and progress the state-of-the-art. Multilingual models and cross-lingual transfer learning are promising avenues to explore for these languages.

3. Dataset and Task Description

3.1. Tasks Description:

Task 1 of the HASOC'23 aimed at identifying hate, offensive, and profane content in social media posts in two languages, namely Sinhala(Task 1a) and Gujarati(Task 1b) [19].

Task 4 was similar to Task 1 and required us to detect hate speech in three other Indian languages, Bengali, Assamese and Bodo. For all the given languages, the training and test datasets had already been provided.

Creating coarse-grained binary classification models to divide tweets into the following two categories was the primary goal for the tasks:

- Hate and Offensive(HOF): Posts that contain hate speech, vulgarity, or offensive material.
- Non-Hate and Offensive (NOT): Posts devoid of offensive language, hate speech, or any other negative elements.

3.2. Data Source

3.2.1. Sinhala Dataset (Task 1a)

The Sinhala dataset provided for train and test had been sourced from the recently released SOLD: Sinhala Offensive Language Detection dataset, which served as a comprehensive resource for the particular task. The training dataset had been further divided into three columns. The first one consisted of the post id, the second of the tweet text, while the third column consisted of the labels, HOF and NOT, for each of the corresponding tweets in the same row.

3.2.2. Gujarati Dataset(Task 1b)

For Task 1b, the training dataset had 200 tweets, primarily categorized into two labels, HOF and NOT, besides three other columns, including the tweet id, the UserName and the date of creation. It is noteworthy that the exact source of the dataset has not been mentioned in the materials provided for the competition.

3.2.3. Assamese, Bengali and Bodo Dataset(Task 4)

The training and test datasets for the task had already been provided in all three languages, Assamese Bengali and Bodo. However, it is worth noting that none of the sources for the data were explicitly mentioned.

3.3. Data Statistics

Distribution of HASOC'23 training datasets for Task 1 and Task 4. For each language, the total no of text entries and the corresponding no of tweets per class are shown below.

Table 1

Task 1 & 4 dataset statistics.

Language	Total text entries	Hate and Offensive(HOT)	Not Hate(NOT)
Assamese	4035	2346	1689
Bengali	2180	515	1665
Bodo	1678	998	680
Sinhala	7500	3176	4324
Gujarati	200	100	100

Besides this, any external use of data beyond what was provided required explicit permission from the competition organizers.

Test Set size:

For Task 1A, the Sinhala test dataset consisted of 2500 tweets. This had to be labeled as either HOF or NOT based on our model.

For Task 1B, the Gujarati test dataset consisted of a total of 1196 tweets to be labelled similarly.

For Task 4, the Bengali, Assamese, and Bodo test datasets consisted of 320, 1009, and 420 text entries, respectively. These entries had to be labeled as either HOF or NOT based on our model.

3.4. Data Preprocessing

We used a number of standard preprocessing methods prior to training our model using the given datasets. Given that the training datasets provided had their texts sourced from Twitter, it was anticipated to contain certain unwanted elements, such as emojis, URLs, mentions and special characters. In order to guarantee the accuracy and relevancy of the text data, we followed procedures to remove such unwanted noise.

4. Methodology

In this section, we describe the methodology and the experimental setup used for the various tasks under HASOC 2023. We conducted a thorough investigation into various neural network architectures, pretrained Large Language models, and classical machine learning models to identify the most effective model for the task.

4.1. Task 1: Identifying Hate, offensive and profane content in Sinhala & Gujarati

For the task of Identifying Hate, offensive and profane content in Sinhala & Gujarati, the models that resulted in the best performance are as follows:

- A CNN-based Binary Classification Model with FastText Embeddings.
- A CNN-BiLSTM based Hybrid Model with FastText/GloVE Embeddings

4.1.1. CNN + FastText Binary Classification Model :

Inspired by the works of Kim et al., [20], we developed the model based on the CNN architecture. At the core of our model lies the input layer, where text sequences representing individual posts are processed. To prepare the input data, we concatenate the words within each sentence, with the sequence length capped at 70 words. The words here are represented as dense vectors of 300 dimensions using pre-trained FastText embeddings for the respective languages. Using machine learning or related dimensional reduction techniques, word embedding converts each token into a vector of real numbers to quantify and classify the semantic similarity of linguistic phrases based on their distributional qualities in a large corpus.

For the convolutional layer, we employed a one-dimensional convolution operation utilizing 100 filters with a kernel size of 3, leading to a systematic scanning of the text sequences and identifying pertinent patterns in the data. An activation function, the rectified linear unit (ReLU), was also applied to introduce non-linearity and enable the model to capture complex relationships in the data. Subsequently, a dense layer with 50 neurons and a ReLU activation function, coupled with an L2-norm constraint, was added to transform the extracted features further. Dropout with a rate of 0.5 was applied as a regularization technique to prevent

overfitting. The resultant vector was then concatenated with the feature vector and the output was passed onto a dense output layer with sigmoid activation and cross entropy loss as shown in Figure 1, to produce the binary hate classification for the model.

4.1.2. CNN-BiLSTM + FastText/GLoVE Binary Classification Model :

Based on the contributions of Vetagiri et al., [4], we developed the model, which is a combination of two different model architectures - the Convolutional Neural Networks (CNN) (Kim, 2014)[20] layer for identifying local textual patterns in the input text and Bidirectional Long Short-Term Memory (BiLSTM) Liu and Guo et al., [21] layer as a form of the Recurrent Neural Architecture Sherstinsky et al., [22] for understanding the long-term complex sequential dependencies within the text data.

The output of these two layers is then passed through a dense layer with a sigmoid activation function for Binary Classification. To prepare the Input data, the model uses the exact pre-trained FastText embeddings for the respective language mentioned above, representing the words as 300-dimensional dense vectors, with the sequence length capped at 70 words, which was held as non-trainable. A similar implementation of this model using pre-trained GloVe embeddings Kumar et al., [23] showed identical results. For the CNN layers, we first employed a SpatialDropout1D layer, a dropout variant that selectively drops entire 1D feature maps during training, to combat overfitting. Subsequently, a one-dimensional convolution layer with 64 filters and a kernel size of 3 was used to capture local textual patterns with fine granularity.

For the BiLSTM part, we used the initial layer with 128 units and a return sequence setting with a dropout of 0.1 and recurrent dropout of 0.1 to process the text inputs in both forward as well as reverse directions followed by a Global Average Pooling and a dense layer with 128 units and a rectified linear unit (ReLU) activation function to introduce non-linearity. Subsequently, a dropout layer is employed whose output is then concatenated with the feature vector and passed through a dense layer with sigmoid activation as shown in figure 2 to produce the overall model for Binary Hate Classification. Our models are trained using the RMSprop optimiser, and our loss function is a binary cross-entropy function. To fine-tune our hyper-parameters over a range of values, we conduct a grid search and select the best-performing model according to validity accuracy. However, no attempt at experimentation by reversing the order of the CNN & Bi-LSTM Layers was made for this particular task.

4.2. Task 4: Identifying Hate, offensive and profane content in Bengali, Bodo, and Assamese languages

The two model architectures used in Task 1 were also experimented with in Task 4. These architectures were used to implement the CNN and the CNN-BiLSTM models which used pre-trained FastText embeddings in the respective languages (except for Bodo, for which Hindi embeddings were used) representing each word as a dense vector with 300 dimensions.

In addition to these two architectures, several others were also experimented with, the details for which are discussed below:

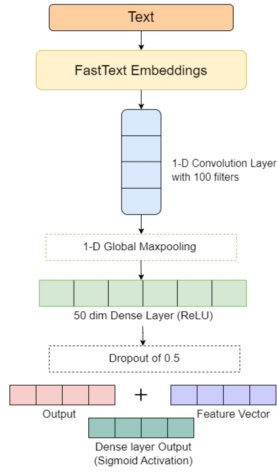


Figure 1: CNN + FastText Binary Classification Model Architecture

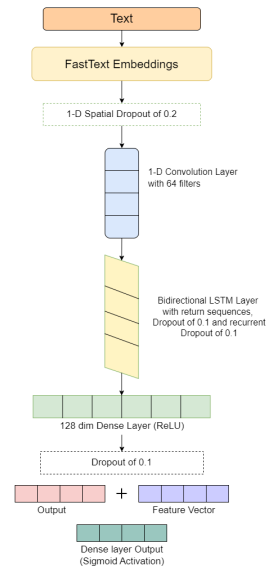


Figure 2: CNN-BiLSTM + FastText/GLoVe Binary Classification Model Architecture

4.2.1. Pre-trained BERT Architecture

Considering the low-resource nature of Task 4 and the limited size of the datasets, we experimented with pre-trained models based on the Bidirectional Encoder Representations from Transformers (BERT) architecture Devlin et al., [24] [25]. We experimented with the Tensorflow Hub to access the pre-trained BERT models.

We used the "bert-multi-cased-preprocess/3" for text processing and the "bert-multi-cased-L-12-H-768-A-12/4" encoder for contextualized word embeddings from the TensorFlow Hub, which is trained on multilingual Wikipedia Data. The model utilizes a BERT preprocessing layer for tokenization and embeddings of the input text, followed by a BERT encoder layer to generate contextual embeddings containing the complex contextual relationship within the language. This is followed by a dropout Neural Layer with a 10% dropout rate to enhance generalization and mitigate overfitting. The final trainable dense layer with 769 employs the sigmoid activation function, producing the binary classification outputs for the given languages

4.2.2. GPT-2 Model

We also explored GPT-2 as a state-of-the-art pre-trained large language model for the task. GPT-2 is a transformer-based model that takes a sequence of words, represented as dense vectors, as input and uses many intermediate layers to extract contextual information for the input text. The output is then passed through a dense layer, producing the Binary Classifier. For the task, we used a pre-trained GPT-2 that contained 768 parameters, fine-tuned on the training dataset for each language in an 80-20% split and a further 20% split from the training set for validation. The input text is tokenised and passed through the model for fine-tuning.

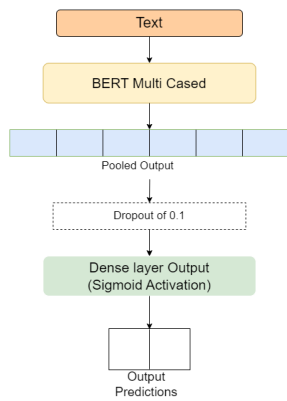


Figure 3: Pre-trained BERT Model Architecture

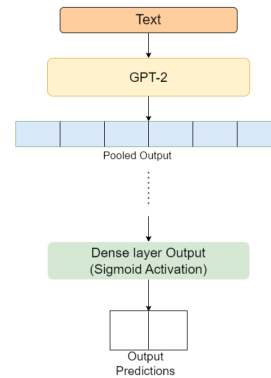


Figure 4: GPT-2 Model Architecture

The model uses an Adam optimizer for optimization with a learning rate of $1e-5$. The batch size for the model is 8

4.2.3. Classical Machine-Learning Based Models

Due to the small size of the training datasets provided for all three languages, we tried implementing classical Machine-Learning approaches as well to accomplish our goals. We experimented with multiple Machine-Learning architectures such as Support Vector Machines (SVMs), Linear Regression, Logistic Regression and Random Forest Classifiers.

However, although the training set accuracies and the macro F1 scores for all the above architectures were quite close, we observed that the Logistic Regression Model achieved the best overall performance. Hence, we created a simple Logistic Regression Model using SciKitLearn Library, trained the Model with the training datasets for each language, and implemented simple tf-idf vectorisation for embedding the input sentences

5. Results and Analysis

5.1. Task 1 & 4

For evaluation the models, the test accuracy for this test data was used for the initial evaluation of the models. In addition to this, the Macro F1 scores, which were acquired by the models on submission in the HASOC 2023 Submission portal were also considered. For Task 4, although the performance of each of the aforementioned models was analyzed, only a few of the models for each language gave the best performance, summarized in Tables 1 & 2.

As it is evident from Tables 1 & 2, the CNN-BiLSTM+FastText/GLoVe Model gave the best Macro F1 Score for Task 1A(Sinhala), and the CNN+FastText Model gave the best Macro F1

Table 2

Performance scores for Tasks 1A & 1B in terms of their test accuracies and Macro F1 scores

Task	Language	Model	Accuracy	Macro F1
1A	Sinhala	CNN+FastText	0.7800	0.7556
		CNN-BiLSTM+FastText	0.7781	0.7711
1B	Gujarati	CNN+FastText	0.7025	0.6873
		CNN-BiLSTM+FastText	0.7121	0.6758

Table 3

Performance scores for Task 4 in terms of their test accuracies

Language	Model	Accuracy	Macro F1
Bengali	CNN+FastText	0.6381	0.60108
	CNN-BiLSTM+FastText/GLoVe	0.6122	
	CNN+FastText+ External Dataset	0.8177	
	GPT-2	0.5924	
	BERT	0.5992	
	Logistic Regression	0.6325	
Assamese	CNN+FastText	0.6374	0.59485
	CNN-BiLSTM+FastText/GLoVe	0.6183	
	GPT-2	0.5825	
	Logistic Regression	0.6559	
Bodo	CNN+FastText	0.6577	
	CNN-BiLSTM+FastText/GLoVe	0.6162	
	Logistic Regression	0.6755	0.66925

Score for Task 1B(Gujarati). For Task 4, in both Bengali and Assamese, the CNN+FastText Model gave the highest accuracy with a Macro F1 score of 0.60108 in Bengali and 0.59485 in Assamese. For Bodo, evidently, the simple Logistic Regression Model gave a much better performance than the other model architectures with a Macro F1 score of 0.66925, possibly due to the small size of the training dataset.

6. Conclusion and Future Scope

Our research has been dedicated to the vital task of identifying hate speech, particularly in Indo-Aryan languages such as Bengali, Assamese, Bodo, Gujarati, and Sinhala. We've devised a comprehensive strategy that unites these linguistic intricacies under a single versatile model. Through domain-aware pre-training and meticulous alignment of our models with language-specific context, we've significantly enhanced hate speech detection. Furthermore, our exploration into model ensemble techniques has bolstered detection accuracy and resilience across diverse language settings, laying a foundational step towards comprehensive hate speech detection in Indo-Aryan languages. Our overarching goal is to foster a safer and more inclusive digital space for speakers of diverse linguistic backgrounds.

As we look to the future, our work paves the way for further advancements in hate-span

detection, focusing on model efficiency, interpretability, and an expansive training data corpus encompassing evolving hate speech trends and linguistic variations. We also recognize the potential of real-time monitoring and context-aware integration in dynamically evolving online environments.

Acknowledgments

We wish to extend our appreciation to the Computer Science and Engineering Department of the National Institute of Technology Silchar for granting us the opportunity to carry out our research and experiments. We are grateful for the support, resources, and research environment offered by the CNLP & AI Lab at NIT Silchar.

References

- [1] B. Vidgen, L. Derczynski, (2020), Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one* 15 (2020).
- [2] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* 51 (2018) 1–30.
- [3] S. Satapara, S. Masud, H. Madhu, M. A. Khan, M. S. Akhtar, T. Chakraborty, S. Modha, T. Mandl, Overview of the HASOC subtracks at FIRE 2023: Detection of hate spans and conversational hate-speech, in: *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.*
- [4] A. Vetagiri, P. Adhikary, P. Pakray, A. Das, CNLP-NITS at SemEval-2023 task 10: Online sexism prediction, PREDHATE!, in: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 815–822. URL: <https://aclanthology.org/2023.semeval-1.113>. doi:10.18653/v1/2023.semeval-1.113.*
- [5] D. Jurgens, L. Hemphill, E. Chandrasekharan, A just and comprehensive strategy for using NLP to address online abuse, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.*
- [6] 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.*
- [7] 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.*
- [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] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 14867–14875.
- [10] S. Munasinghe, U. Thayasivam, A deep learning ensemble hate speech detection approach for sinhala tweets, in: 2022 Moratuwa Engineering Research Conference (MERCon), 2022, pp. 1–6. doi:10.1109/MERCon55799.2022.9906232.
- [11] 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, 2019, pp. 1–8. doi:10.1109/ICTer48817.2019.9023655.
- [12] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, M. Zampieri, Sold: Sinhala offensive language dataset, 2022. arXiv:2212.00851.
- [13] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media, 2018, pp. 36–41.
- [14] R. Khurana, C. Pandey, P. Gupta, P. Nagrath, AniMOJity: detecting hate comments in Indic languages and analysing bias against content creators, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON), Association for Computational Linguistics, New Delhi, India, 2022, pp. 172–182. URL: <https://aclanthology.org/2022.icon-main.23>.
- [15] A. Das, N. Tandon, S. Narayan, HASoC: Hate Speech and Offensive Content Identification in Indo-European Languages: Overview of HASoC Track at FIRE 2020, in: Forum for Information Retrieval Evaluation (FIRE), 2020, pp. 29–32.
- [16] 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.
- [17] B. R. Chakravarthi, R. Priyadarshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 133–145. URL: <https://aclanthology.org/2021.dravidianlangtech-1.17>.
- [18] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, “Leveraging GPT-2 for Automated Classification of Online Sexist Content“, In Exist 2023 Lab at CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece, 2023.
- [19] 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.
- [20] Y. Kim, Convolutional neural networks for sentence classification, 2014. arXiv:1408.5882.
- [21] G. Liu, J. Guo, Bidirectional lstm with attention mechanism and convolutional

- layer for text classification, *Neurocomputing* 337 (2019) 325–338. URL: <https://www.sciencedirect.com/science/article/pii/S0925231219301067>. doi:<https://doi.org/10.1016/j.neucom.2019.01.078>.
- [22] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, *Physica D: Nonlinear Phenomena* 404 (2020) 132306. URL: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>. doi:<https://doi.org/10.1016/j.physd.2019.132306>.
- [23] S. Kumar, S. Kumar, D. Kanojia, P. Bhattacharyya, “a passage to India”: Pre-trained word embeddings for Indian languages, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 352–357. URL: <https://aclanthology.org/2020.sltu-1.49>.
- [24] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR* abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [25] 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>.