# Multilingual Hate Speech Detection Using Ensemble of Transformer Models

Md Saroar Jahan[1,*,†], Fadi Hassan[1,†], Walid Aransa[1,†] and Abdessalam Bouchekif[1,†]

[1]*Huawei Finland Research Center*

Abstract

The classification of hate speech and offensive language presents significant challenges, primarily due to the scarcity of low-resource datasets and the absence of pre-trained models. This paper offers a comprehensive overview of offensive language identification results in the context of HASOC-2023 across various languages and tasks, including Sinhala and Gujarati, Bengali, Assamese, and Bodo, and Hateful span detection. To address these challenges, we harnessed the power of BERT-based models, leveraging resources such as XLM-RoBERTa-large, l3-cube, BanglaHateBert, and BenglaBERT. Our research findings yielded promising results, notably showcasing the superior performance of XLM-RoBERTa-large over monolingual models in the majority of cases. For Task 3, SpanBERT performed outstandingly.

Notably, our team FiRC-NLP contributions were acknowledged with top-ranking achievements, securing the first position in Task 1, and Task 3, while clinching the second position in Task 4.

**Keywords**

Hateful Span Detection, Conversational Hate Detection, SpanBERT

## 1. Introduction

Social media is a widely popular and convenient platform for open expression and online communication with others. Unfortunately, it also provides the means for distributing abusive and aggressive content such as sexism, racism, politics, cyberbullying, and blackmailing. Nockleby [1] stated that "hate speech disparages a person or group based on some characteristics such as race, color, and ethnicity". Addressing offensive language on social media is now a major challenge. Various shared tasks and data-sharing initiatives within the research community aim to motivate researchers to develop innovative solutions for detecting abusive content. Among the initiatives, HASOC has gained significant popularity, with its previous editions: HASOC-2019 [2], HASOC-2020[3], HASOC-2021[4] and HASOC-2022[5]. These editions focus on Hate speech and offensive language identification in English, German, and Hindi. SemEval is another noteworthy initiative. SemEval-2019 [6] focuses on the detection of hate speech against immigrants and women in Spanish and English messages extracted from Twitter. SemEval-2020 [7] extends its scope to include Arabic, English, Danish, Greek, and Turkish content. In SemEval-2023 [8], the focus is on detecting and identifying comments and tweets containing

*Forum for Information Retrieval Evaluation, December 15–18, 2023, India*

*Corresponding author.

†These authors contributed equally.

saroar.jahan@huawei.com (M. S. Jahan); fadi.hassan@huawei.com (F. Hassan); walid.aransa@huawei.com (W. Aransa); abdessalam.bouchekif@huawei.com (A. Bouchekif)

sexist expressions. Additionally, other shared tasks are proposed, such as GermEval [9] for the German language, EVALITA[10] for Italian languages, and OSACT [11] for Arabic content, all of which contribute to this important area of research."

The best models developed for including models such as Roberta [12], DeBERTa [13], ALBERT [14] and XLM-RoBERTa [15]. In SemEval 2023 task 10-A, the top-performing models [16] and [17] are based on DeBERTa. The best performing models in SemEval 2022 task 12-A [18] used an ensemble of ALBERT models of different sizes, while the second-ranked team [19] used Roberta-base and XLM-Roberta. Monolingual Transformers give better results when addressing challenges related to low-resource languages, compared to Multilingual. The winning team in SemEval-2020 Task 12-A for Arabic [20] and Danish [21] languages achieved highest performance by using AraBERT [22] and Nordic BERT [1], respectively.
Hate speech detection becomes more challenging when social posts are written in a Code-Mixed (CM) language. Code-mixing, the practice of blending words from two languages within a single sentence, is becoming increasingly common in various bilingual communities, which renders the automatic making detection task more challenging [23, 24]. In HASOC-2022, three tasks were hosted: Task 1 and Task 2 involved binary and multi-class classification for both German and code-mixed languages, while Task 3 focused on identifying offensive language in Marathi. The highest performance in Task 1 [25] was achieved using Google-MuRIL [2] (BERT model pre-trained on 17 Indian languages). HASOC 2023 introduced Task 1 and Task 4, focusing on the detection of hate speech, offensive content, and profanity. Task 3 centered on detecting hate speech spans within social media posts. We actively engaged in this competition, taking on Task 1 for languages including Bengali, Gujarati, Sinhala, Assamese, and Bodo, additionally, we took on Task 3 which involved English hate speech span detection. To accomplish these tasks, we made use of the HASOC-2023 shared dataset for both training and validation purposes without any external data.

Our strategy predominantly relied on cutting-edge transformer models to tackle these challenges. This paper is structured as follows: Section 2 presents a detailed description of the tasks and datasets, Section 3, we provides an in-depth look at our methodology and model architecture. Lastly, the conclusion section offers definitive statements and delineates potential directions for future research.

## 2. Task Description

This section presents the task descriptions for HASOC 2023 [26] as follows:

**Task 1 and 4** focus on identifying hate speech, offensive language, and profanity in different languages using natural language processing techniques [27, 28, 29, 30, 31, 32, 33]. These task mainly involves classifying tweets into two categories: Hate and Offensive (HOF) or Non-Hate and Offensive (NOT).

- Task 1A: deals with identifying hate and offensive content in Sinhala, a low-resource Indo-Aryan language spoken in Sri Lanka.

[1]https://github.com/certainlyio/nordic_bert
[2]https://huggingface.co/google/muril-base-cased

- Task 1B: focuses on identifying hate and offensive content in Gujarati, another low-resource Indo-Aryan language spoken by approximately 50 million people in India. The training set for this task consists of around 200 tweets.
- Task 4: aims to detect hate speech in Bengali, Bodo, and Assamese languages. Data is primarily collected from Twitter, Facebook, or YouTube comments.

**Task 3** aims to detect the various hateful spans within a sentence already considered hateful [34]. The input texts are all in English. The detection of hateful spans is achieved by mapping this into a sequence labeling problem. For every token of the sequences, human annotators have manually annotated the start and end of a hateful span. This is achieved by the BIO notation tagging, where 'B' represents the beginning of the hate span,' I' forms the continuation of a hate span, and 'O' represents the non-hate tag.

**Table 1**
Example of datasets of task 1 and 4

| **Sentence** | **Translation** | **Label** | **Task, Language** | **Train and Test size** |
|---|---|---|---|---|
| "বালের শিক্ষা মন্ত্রী" | Stupid Education Minister | HOF | Task4, Bengali | 1281, 320 |
| "কুকুৰ বুলি কিয় কৈছে অসভ্য ক'ৰবাৰ, লাজ নাই" | Why are you calling me a dog, rude somewhere, no shame | HOF | Task4, Asamee | 4036, 1009 |
| "मोसौ खुगायाव एमफौ नांबाय नोंनाव सैम" | Both are drunkards f***rs | HOF | Task 4 Bodo | 1679, 420 |

# 3. Methodology

This section offers a comprehensive overview encompassing the model architecture description and the strategies employed to address each task. Due to the similarities between Task 1 and Task 4, we have consolidated them into a single section, while Task 3 are separately described.

## 3.1. Task 1 and 4 Model Architecture

For Task 1 and Task 4, we adopted two main strategies:

1. Utilizing Different BERT Models: We conducted experiments with both multilingual and monolingual BERT models.
2. Augmenting Training Data: Our second approach involved enhancing the training data through automatic annotation.

We assessed several models, including multilingual ones such as XLM-RoBERTa-large and IndicBERT, and monolingual models like L3-cube, Bangla BERT, and Bangla Hate BERT.

Following our experiments, we selected XLM-RoBERTa-large as the baseline model due to its superior performance when compared to all the monolingual models. This performance difference may be attributed to the age of some monolingual models, such as Bangla Hate BERT, which is considerably older compared to XLM-RoBERTa-large. However, for the Bangla L3-cube monolingual model, it exhibited a slightly better performance by +0.06 F1 score compared to XLM-RoBERTa-large. Nevertheless, since XLM-RoBERTa-large outperformed most monolingual models in most cases, we opted to choose it as the baseline model.

**Table 2**
Task 1 F1 Scores for Different Models by Language

| Language | Model | F1 Score |
|---|---|---|
| Gujarati | Indic-Bert (trained on 12 major Indian languages) | 73.4 |
| | L3-cube Gujarati (monolingual) | 79.1 |
| | XLM-RoBERTa-large | 81.6 |
| Sinhala | Indic-Bert (trained on 12 major Indian languages) | 74.5 |
| | L3-cube Sinhala (monolingual) | 78.6 |
| | XLM-RoBERTa-large | 80.4 |
| Bengali | Indic-Bert (trained on 12 major Indian languages) | 70.5 |
| | L3-cube Sinhala (monolingual) | 75.6 |
| | XLM-RoBERTa-large | 75.1 |
| | banglaBERT | 68.1 |
| | BanglaHateBERT | 65.5 |

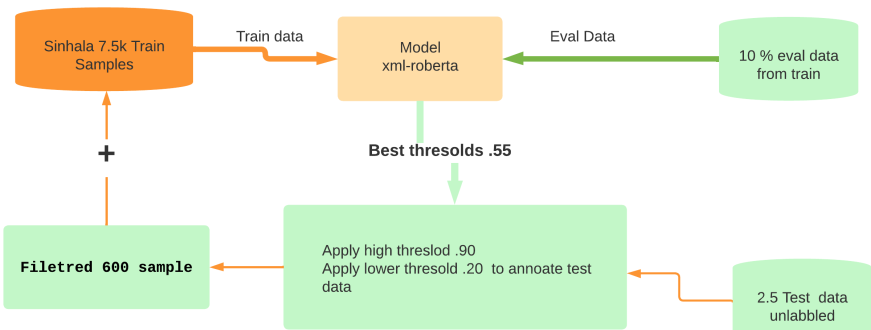


**Figure 1:** Illustration of the Approach to Enhance Model Performance: Incorporating Annotated Test Data into Training Data for Sinhala (Similar approach tested with public data).

To further enhance our model's performance, we pursued a second strategy, which involved expanding our training dataset. However, due to a lack of suitable datasets for most of these languages, we hypothesized that incorporating automatically annotated test data into the training data could improve model learning. To implement this, we initially trained the model using 90% of the training data and 10% of the evaluation data. We then determined the optimal thresholds during evaluation and applied upper and lower thresholds to automatically annotate part of the test data. For example, we used a 0.90 upper threshold and a 0.20 lower threshold. After automatically annotating that portion of test data with these thresholds, we retrained the model by adding this part of test data to the training data and observed a 3% improvement in model performance. This hypothesis was further tested with external public data, where automatic annotations were applied using these upper and lower thresholds, resulting in a 1-2% improvement. Additionally, we also employed the ensemble of 5 models, which contributed to a 0.4% increase in F1 scores (see Table 3).

**Table 3**
F1 Scores for Different Models by Language after Adding part of Test Data to Training and Using Ensemble Model

| Language | Model | F1 Score |
|---|---|---|
| Gujarati | XLM-RoBERTa-large | 81.6 |
| | XLM-RoBERTa-large 5 ensemble model | 82.0 |
| | XLM-RoBERTa-large (200 train data + adding filtered test data 401 sample) | 84.8 |
| Sinhala | XLM-RoBERTa-large | 80.4 |
| | XLM-RoBERTa-large 5 ensemble model | 80.9 |
| | XLM-RoBERTa-large (7.5k train data + adding filtered test data 600 sample) | 83.8 |

## 3.2. Task 3 Model Architecture

In Task 3, the goal is to find all the hateful spans. A hateful span is a group of words that together express the hatred in the sentence. In this task, the provided data size: 1936 training samples, 485 validation samples, and 606 test samples, and the specific labels used for this task, such as "B-HateSpan" to denote the first token in a hateful span and "I-HateSpan" to indicate tokens inside a hateful span. Additionally, the section includes an analysis of the model's performance and an in-depth explanation of the outcomes.

The architecture of our best submitted model for Task 3 employs a teacher-student framework and utilizes the SpanBERT-base-cased model along with Conditional Random Fields (CRF) for sequence tagging. The approach can be summarized as follows:

- Teacher Model: Ensemble of $k$ SpanBERT-base-cased models, each combined with CRF.
- Student Model: A single SpanBERT-base-cased model, also integrated with CRF. The student model is distilled from the teacher model using a specific formula:

$$\mathcal{L}_{\text{loss}} = (1 - \alpha) \cdot \text{CE}(student\_score, target) + \alpha \cdot \text{MSE}(student\_logits, teacher\_logits) \quad (1)$$

1. **Model Comparison:** Table 5 provides a comparison of different base models with varying configurations, including casing, k-fold cross-validation, and tagging schemes (BIO and IO). It is observed that SpanBERT-large with lower casing and a 5-fold cross-validation scheme achieved the highest private score of 62.322, indicating its effectiveness in identifying hateful spans.
2. **Impact of Casing:** The casing of the model input, whether lower case or true case, seems to affect the model's performance. Lower casing generally performs better, as indicated by the higher private and public scores in several configurations.

**Table 4**
Evaluation of hate span detection performance utilizing various models, with the submitted model highlighted in bold.

| Base Model | Casing | K-fold | Tagging | Private Score | Public Score |
|---|---|---|---|---|---|
| SpanBERT-large | Lower case | 5 | BIO | 62.322 | 55.052 |
| SpanBERT-base | True case | - | BIO | 41.528 | 33.755 |
| SpanBERT-base | True case | 5 | BIO | 55.547 | 48.566 |
| SpanBERT-base | Lower case | 10 | BIO | 57.541 | 51.013 |
| **SpanBERT-base** | Lower case | 5 | BIO | **57.605** | **53.378** |
| SpanBERT-base | True case | - | BIO | 55.177 | 45.602 |
| DeBERTa-v3-xlarge | True case | - | BIO | 43.102 | 38.249 |
| DeBERTa-v3-large | True case | - | BIO | 47.433 | 39.222 |
| DeBERTa-v3-large | True case | - | IO | 15.426 | 12.446 |

3. **Tagging Scheme:** The choice of tagging scheme (BIO vs. IO) also influences performance. Models using the BIO tagging scheme tend to yield better results, as seen in higher private and public scores.
4. **Ensemble vs. Single Model:** The ensemble approach using multiple SpanBERT-base-cased models as teachers seems to provide valuable knowledge transfer to the student model, resulting in improved performance.
5. **Distillation Effect:** The use of distillation with an $\alpha$ value of 0.95 for transferring knowledge from teachers to the student model helps enhance performance compared to a standalone student model (See Eq.1).

Overall, the model architecture involving an ensemble of SpanBERT models with CRF, especially when using lower casing and BIO tagging, demonstrates strong performance in identifying hateful spans in text. The distillation process further boosts the student model's effectiveness.

**Table 5**
The official outcomes from our participation in the HASOC-23 encompassing Task 1, 3, and 4, best models are presented

| Team name | Task, Language | Base model | Macro F1 | Rank |
|---|---|---|---|---|
| FiRC-NLP | Task 1b (Gujrate) | XLM-RoBERTa-large | 0.848 | 1/17 |
| | Task 1a (Sinhala) | XLM-RoBERTa-large | 0.838 | 1/16 |
| | Task 3 (English) | SpanBERT-base | 0.570 | 1/12 |
| | Task 4 (Bengali) | XLM-RoBERTa-large | 0.764 | 2/20 |
| | Task 4 (Assamese) | XLM-RoBERTa-large | 0.725 | 2/20 |
| | Task 4 (Bodo) | XLM-RoBERTa-large | 0.848 | 4/19 |

## 4. Conclusion

In this paper, we have presented a comprehensive analysis of hate speech and offensive language identification across multiple languages and tasks in HASOC-2023 competition. In Task 1 and Task 4, our research involves identifying offensive language in Sinhala, Gujarati, Bengali, Assamese, and Bodo languages, and Task 3, which involves hateful span detection in English text.

Our research not only showcased the effectiveness of transformer-based models in these shared tasks but also emphasized the importance of model selection, task-specific customization, and innovative strategies to address the challenges posed by low resource languages, multilingual and cross-lingual contexts.

As future work, further investigations are needed to explore the use of more diverse and specialized transformer models, as well as fine-tuning model parameters to achieve even better results. Additionally, we need to inspect the application of ensemble techniques and the incorporation of multiple thresholds for automatic annotation represents promising avenues for improving model robustness and generalization.

## References


[1] J. T. Nockleby, Hate speech, Encyclopedia of the American constitution 3 (2000) 1277–1279.

[2] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th forum for information retrieval evaluation, 2019, pp. 14–17.

[3] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, 2020, pp. 29–32.

[4] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021, ACM, 2021, pp. 1–3.

[5] S. Modha, T. Mandl, P. Majumder, S. Satapara, T. Patel, H. Madhu, Overview of the HASOC subtrack at FIRE 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, volume 3395 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 475–488.

[6] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), arXiv preprint arXiv:1903.08983 (2019).

[7] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive

language identification in social media (offenseval 2020), arXiv preprint arXiv:2006.07235 (2020).
[8] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023, Association for Computational Linguistics, 2023, pp. 2193–2210.
[9] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language (2018).
[10] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the evalita 2018 hate speech detection task, in: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 2263, CEUR, 2018, pp. 1–9.
[11] H. Mubarak, H. Al-Khalifa, A. Al-Thubaity, Overview of OSACT5 shared task on Arabic offensive language and hate speech detection, in: Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, European Language Resources Association, 2022, pp. 162–166.
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019).
[13] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, CoRR abs/2111.09543 (2021).
[14] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, CoRR abs/1909.11942 (2019).
[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019).
[16] M. Zhou, PingAnLifeInsurance at SemEval-2023 task 10: Using multi-task learning to better detect online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023.
[17] F. Hassan, A. Bouchekif, W. Aransa, Firc at semeval-2023 task 10: Fine-grained classification of online sexism content using deberta, in: Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023, Association for Computational Linguistics, 2023, pp. 1824–1832.
[18] G. Wiedemann, S. M. Yimam, C. Biemann, UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020.
[19] S. Wang, J. Liu, X. Ouyang, Y. Sun, Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020.
[20] H. Alami, S. Ouatik El Alaoui, A. Benlahbib, N. En-nahnahi, LISAC FSDM-USMBA team at SemEval-2020 task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020.
[21] M. Pàmies, E. Öhman, K. Kajava, J. Tiedemann, LT@Helsinki at SemEval-2020 task 12:

Multilingual or language-specific BERT?, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020.
[22] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 9–15. URL: https://aclanthology.org/2020.osact-1.2.
[23] I. A. Bhat, V. Mujadia, A. Tammewar, R. A. Bhat, M. Shrivastava, Iiit-h system submission for fire2014 shared task on transliterated search, in: Proceedings of the Forum for Information Retrieval Evaluation, 2014, pp. 48–53.
[24] M. L. Ripoll, F. Hassan, J. Attieh, G. Collell, A. Bouchekif, Multi-lingual contextual hate speech detection using transformer-based ensembles, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.
[25] N. K. Singh, U. Garain, An analysis of transformer-based models for code-mixed conversational hate-speech identification, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.
[26] S. Masud, M. Bedi, M. A. Khan, M. S. Akhtar, T. Chakraborty, Proactively reducing the hate intensity of online posts via hate speech normalization, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3524–3534. URL: https://doi.org/10.1145/3534678.3539161. doi:10.1145/3534678.3539161.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.

[33] 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.

[34] S. Masud, M. A. Khan, M. S. Akhtar, T. Chakraborty, Overview of the HASOC Subtrack at FIRE 2023: Identification of Tokens Contributing to Explicit Hate in English by Span Detection, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.