# Overview of the HASOC Subtrack at FIRE 2023: Hate-Speech Identification in Sinhala and Gujarati

Shrey Satapara[1], Hiren Madhu[2], Tharindu Ranasinghe[3], Alphaeus Eric Dmonte[4], Marcos Zampieri[4], Pavan Pandya[5], Nisarg Shah[5], Sandip Modha[6], Prasenjit Majumder[7] and Thomas Mandl[8]

[1]*Indian Institute of Technology, Hyderabad, India*
[2]*Indian Institute of Science, Bangalore, India*
[3]*Aston University, United Kingdom*
[4]*George Mason University, USA*
[5]*Indiana Bloomington University, USA*
[6]*LDRP-ITR, Gandhinagar, India*
[7]*DA-IICT, Gandhinagar, India*
[8]*University of Hildesheim, Germany*

## Abstract

Detecting offensive and hateful content in low-resource languages poses a significant challenge due to the limited availability of benchmark datasets. It is crucial to address this gap by creating benchmark datasets tailored to these languages. This not only enhances the accuracy of detection but also provides valuable insights into the efficacy of identifying problematic content in comparison to high-resource languages. In line with this commitment to advancing research on low-resource languages, the Hate Speech and Offensive Content Identification (HASOC) shared task introduced a dedicated subtrack for Hate Speech Identification in Sinhala and Gujarati in 2023. This paper outlines the objectives of the task, discusses the characteristics of the data involved, and presents an analysis of the participants' submissions. For Task 1a, we utilized an existing Sinhala dataset (SOLD) consisting of 10,000 tweets. Meanwhile, for Task 1b, focused on Gujarati, we curated a new dataset comprising 1,020 tweets. A total of 16 teams submitted experiments for Sinhala, with the leading team achieving an impressive F1 score of 0.83. In the case of the Gujarati task, 17 teams participated, and the highest-performing team achieved an F1 score of 0.84. These results highlight the significance of tailored datasets in facilitating the effective detection of offensive content in low-resource languages.

## Keywords

Hate Speech, Social NLP, Social Media, Language Resource, Deep Learning, Low-Resource Language, Evaluation, Benchmark

*Forum for Information Retrieval Evaluation, December 15-18, 2023, Goa, India*
shreysatapara@gmail.com (S. Satapara); hirenmadhu16@gmail.com (H. Madhu); t.ranasinghe@aston.ac.uk (T. Ranasinghe); admonte@gmu.edu (A. Dmonte); marcos.zampieri@rit.edu (M. Zampieri); pavanpandya1311@gmail.com (P. Pandya); nisarg0606@gmail.com (N. Shah); sjmodha@gmail.com (S. Modha); p_majumder@daiict.ac.in (P. Majumder); mandl@uni-hildesheim.de (T. Mandl)
0000-0001-6222-1288 (S. Satapara); 0000-0002-6701-6782 (H. Madhu); 0000-0002-2346-3847 (M. Zampieri); 0000-0003-2427-2433 (S. Modha); 0000-0002-8398-9699 (T. Mandl)

# 1. Introduction

Hate speech is a global problem that plagues social media platforms in many countries [1]. Hate speech can ultimately also lead to violent hate crimes [2]. Consequently, detection and moderation are necessary to maintain a rational discourse that allows an exchange of arguments. Reduced efforts in content moderation can lead to the proliferation of hate speech, as the case of Twitter has shown recently [3].

The initiative Hate Speech and Offensive Content Identification (HASOC) has organized shared tasks since 2019 [4] and created resources for several languages. The efforts for the creation of language resources for low-resource languages are of special importance. Research needs to analyze which resources are beneficial for such languages. Is it better to develop specific resources, or is it better to use resources from high-resource languages like English and exploit this knowledge for a low-resource context (e.g., by translating content or transfer learning between languages) [5].

Many offensive language detection benchmarks are available for English and other high-resource languages. However, in the last few years, the NLP community has focused on creating more datasets for low-resource languages such as Marathi [6], Oromo [7], Swahili [8] Greek [9], Danish [10], and Albanian [11]. Supporting this, the last two editions of HASOC contained shared tasks on identifying offensive language in Marathi.

In 2023, the HASOC subtask 1 focused on identifying hate speech, offensive language, and profanity in Sinhala and Gujarati. Sinhala is a low-resource Indo-Aryan language spoken by around 16 million people, mainly in Sri Lanka. Gujarati is also a low-resource Indo-Aryan language spoken by approximately 50 million people, mainly in North-Western India.

Task 1A deals with identifying hate and offensive content in Sinhala. The task involves classifying tweets into Hate and Offensive (HOF) or Non-Hate and Offensive (NOT). The dataset for this task is based on the Sinhala Offensive Language Detection dataset (SOLD) [12]. Task 1B focuses on identifying hate and offensive content in Gujarati, which was similar to task 1A, where the participants need to classify tweets into HOF or NOT categories. We created a new dataset for Gujarati with 1020 annotated tweets. More details about the dataset is available in Section 2.

Overall, both tasks were highly successful and gained attention of the NLP community. The interest demonstrated last year continued this year, too, with 16 teams participating in the Sinhala task and 17 teams participating in the Gujarati task. Furthermore, it should be highlighted that this is the first-ever shared task organised for Sinhala. We believe that this shared task would open many research avenues for low-resource languages like Sinhala and Gujarati.

# 2. Data

## 2.1. Sinhala dataset

The data used for subtask 1A are from the Sinhala Offensive Language Dataset: **SOLD**[1] [12]. The dataset consists of 10,000 annotated Twitter posts aimed to detect Sinhala offensive text. The

[1]https://huggingface.co/datasets/sinhala-nlp/SOLD

dataset has two splits, the training and test sets, containing 7500 and 2500 tweets, respectively. The initial dataset consisted of two annotation levels, which were sentence and token-level annotations. They have followed the OLID [13] Task A annotation for sentence level annotations, which we utilized for our subtask 1A. The original paper demonstrates 0.7 - 0.8 Fleiss' Kappa Inter Annotator Agreement to this dataset. Class distribution of the dataset is shown in Table 1.

| Class | Train | Test |
|---|---|---|
| HOF | 3176 | 1015 |
| NOT | 4324 | 1485 |

**Table 1**
Class distribution in SOLD [12]

### 2.2. Gujarati dataset

We created a new Gujarati offensive language detection dataset for subtask 1B. We used the Sinhala dataset from subtask 1A to create the dataset. We first collected all the unique offensive tokens from the Sinhala dataset. These tokens were then automatically translated to Gujarati. From the translations, we manually selected 45 tokens. We also collected offensive tokens from various websites (eg: https://www.youswear.com/index.asp?language=Gujarati ) and manually selected the ones that were appropriate for our problem statement. We then used an in-house web scraper to scrape the tweets using those keywords.

We present the dataset statistics for the Gujarati dataset in Table 2. As we can see, we only provide the participants with 200 labelled text samples. This was done to encourage participants to develop innovative techniques in Zero-Shot and Few-Shot learning that make use of high-resource datasets. For the annotations, the inter-annotator agreement was 0.7474.

| Class | Train | Test |
|---|---|---|
| HOF | 100 | 376 |
| NOT | 100 | 820 |
| Total | 200 | 1196 |

**Table 2**
Class distribution in Gujarati

## 3. Results

The results for Subtask 1A are presented in Table 3. A total of 52 systems were submitted from 16 teams. The best-performing system by each team is displayed in table 3, ranked by their F1 scores. The performance of the top-5 teams was very similar. Most of the top teams utilized pre-trained transformer models that support Sinhala, such as XLM-R. Several teams used sentence embeddings such as SBERT and LABSE in their experiments. Interestingly, some teams used mBERT, which is not trained on Sinhala text but could also achieve mid-table finishes. Team FiRC-NLP had the best performing system, with an F1 score of 0.8382, followed by "Krispy

| Rank | Team Name | Number of Runs | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 | FiRC-NLP[14] | 2 | 0.8368 | 0.8399 | 0.8382 |
| 2 | Krispy Mango[15] | 5 | 0.8439 | 0.8326 | 0.8371 |
| 3 | AiAlchemists[16] | 3 | 0.8339 | 0.8374 | 0.8355 |
| 4 | SATLab[17] | 5 | 0.8377 | 0.833 | 0.8351 |
| 5 | Z-AGI Labs[18] | 4 | 0.8342 | 0.8357 | 0.8349 |
| 6 | NAVICK[19] | 5 | 0.8304 | 0.8262 | 0.8281 |
| 7 | XAG-TUD[20] | 1 | 0.8178 | 0.8093 | 0.8127 |
| 8 | Gradient Descenders | 1 | 0.8087 | 0.8059 | 0.8072 |
| 9 | SSN_CSE_ML_TEAM[21] | 5 | 0.7977 | 0.7923 | 0.7946 |
| 10 | IRLab@IITBHU[22] | 4 | 0.7853 | 0.7849 | 0.7851 |
| 11 | MUCS_3 | 5 | 0.8056 | 0.7753 | 0.7832 |
| 12 | CNLP-NITS-PP[23] | 2 | 0.7716 | 0.7707 | 0.7711 |
| 13 | UINSUSKA-Mandiri[24] | 3 | 0.7393 | 0.7455 | 0.741 |
| 14 | Wunderkinds | 1 | 0.6839 | 0.66 | 0.6628 |
| 15 | Hate Speech Detectives | 1 | 0.6446 | 0.6425 | 0.6433 |
| 16 | LEGEND[25] | 5 | 0.5588 | 0.5572 | 0.5574 |

**Table 3**
Results of Subtask 1A - Sinhala. The best system for each team is reported, ordered from best to least performing system

| Rank | Team | Number of Runs | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 | FiRC-NLP[14] | 2 | 0.8391 | 0.8637 | 0.8487 |
| 2 | SATLab[17] | 5 | 0.8500 | 0.8292 | 0.8382 |
| 3 | Krispy Mango[15] | 5 | 0.7896 | 0.8034 | 0.7956 |
| 4 | AiAlchemists[16] | 4 | 0.7859 | 0.8254 | 0.7926 |
| 5 | XAG-TUD[20] | 2 | 0.7717 | 0.7958 | 0.7799 |
| 6 | SSN_CSE_ML_TEAM[21] | 5 | 0.7675 | 0.8048 | 0.7731 |
| 7 | Z-AGI Labs[18] | 4 | 0.7607 | 0.7970 | 0.7660 |
| 8 | LEGEND[25] | 5 | 0.7711 | 0.7415 | 0.7526 |
| 9 | Wunderkinds | 2 | 0.7292 | 0.7606 | 0.7333 |
| 10 | Sanvadita[26] | 6 | 0.7394 | 0.7776 | 0.7324 |
| 11 | MUCS | 7 | 0.7250 | 0.7572 | 0.7276 |
| 12 | NAVICK[19] | 4 | 0.7038 | 0.7364 | 0.6945 |
| 13 | IRLab@IITBHU[22] | 4 | 0.6915 | 0.7205 | 0.6896 |
| 14 | CNLP-NITS-PP[23] | 1 | 0.6998 | 0.7317 | 0.6873 |
| 15 | UINSUSKA-Mandiri[24] | 3 | 0.6929 | 0.7218 | 0.6675 |
| 16 | Gradient Descenders | 2 | 0.6710 | 0.6626 | 0.6661 |

**Table 4**
Results of Subtask 1B - Gujarati. The best system for each team is reported, ordered from best to least performing system

Mango" and "AiAlchemist", with F1 scores of 0.8371 and 0.8355, respectively. The last-ranked team had an F1 score of 0.5574.

Table 4 presents the results of Subtask 1B. Notably, a total of 54 submissions were received from 17 different teams. The team "FiRC-NLP" achieved the highest F1 score (0.8488) with their

submission named ”no-kfold,” demonstrating excellent precision (0.8392) and recall (0.8638) by fine-tuning XLM-RoBERTa large checkpoint. Following closely, ”SATLab” secured the second position with their submission ”HasocT1bR4,” earning an F1 score of 0.8383, by learning character level ngrams with classical machine learning classifiers. The team ”Krispy Mango” by fine-tuning XLM-RoBERTa, ranked third with an F1 score of 0.7956. The table provides an insightful overview of the competition results, highlighting the strong performance of many teams and their respective submissions.

## 4. Conclusion and Future Work

We presented the results of HASOC 2023 Task 1, which featured datasets in two low-resource Indo-Aryan languages; Sinhala and Gujarati. A total of 16 teams submitted experiments for Sinhala and 17 teams participated for Gujarati. The wide participation in the task allowed us to compare a number of approaches. We observed that the best systems for both languages used pre-trained transformers that support Sinhala and Gujarati, such as XLM-R and mBERT. Furthermore, since Gujarati only contained a limited number of training instances, several teams utilised cross-lingual transfer learning to improve their performance. Despite being low-resource language, top teams produced competitive results that are comparable to high-resource languages.

We plan to extend the task in several ways. First, we plan to organize an offensive spans detection task for these two languages that will improve the explainability of the offensive language detection models. Secondly, we hope to add more Indo-Aryan languages that are less researched in the NLP community. HASOC 2023 is the first-ever shared task organised for Sinhala and one of the few shared tasks organized for Gujarati. However, we believe that in light of HASOC 2023, many shared tasks will be created for these languages in the future, improving the involvement of NLP researchers in these low-resource languages.

## Acknowledgments

This work was partially supported by a grant from the Artificial Intelligence Journal (AIJ) for sponsoring IA research (28th call for sponsorship).

## References

[1] B. Di Fátima, Hate Speech on Social Media: A Global Approach, LabCom Books & EdiPUCE, Covilhã, Portugal, 2023. doi:10.25768/654-916-9.

[2] K. Müller, C. Schwarz, From hashtag to hate crime: Twitter and antiminority sentiment, American Economic Journal: Applied Economics 15 (2023) 270–312. doi:10.1257/app.20210211.

[3] D. Hickey, M. Schmitz, D. Fessler, P. E. Smaldino, G. Muric, K. Burghardt, Auditing Elon Musk’s impact on hate speech and bots, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1133–1137. doi:10.1609/icwsm.v17i1.22222.

[4] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019, ACM, 2019, pp. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[5] M. U. Arshad, R. Ali, M. O. Beg, W. Shahzad, Uhated: hate speech detection in urdu language using transfer learning, Lang. Resour. Evaluation 57 (2023) 713–732. URL: https://doi.org/10.1007/s10579-023-09642-7. doi:10.1007/s10579-023-09642-7.

[6] S. S. Gaikwad, T. Ranasinghe, M. Zampieri, C. Homan, Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi, in: G. Angelova, M. Kunilovskaya, R. Mitkov, I. Nikolova-Koleva (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021, INCOMA Ltd., 2021, pp. 437–443. URL: https://aclanthology.org/2021.ranlp-1.50.

[7] N. B. Defersha, J. Abawajy, K. Kekeba, Deep learning based multilabel hateful speech text comments recognition and classification model for resource scarce ethiopian language: The case of afaan oromo, in: IEEE International Conference on Current Development in Engineering and Technology (CCET), IEEE, 2022, pp. 1–11. doi:10.1109/CCET56606.2022.10080837.

[8] E. Ombui, L. Muchemi, P. Wagacha, Building and annotating a codeswitched hate speech corpora, Int. J. Inf. Technol. Comput. Sci 3 (2021) 33–52.

[9] Z. Pitenis, M. Zampieri, T. Ranasinghe, Offensive Language Identification in Greek, in: Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 5113–5119. URL: https://aclanthology.org/2020.lrec-1.629/.

[10] G. I. Sigurbergsson, L. Derczynski, Offensive Language and Hate Speech Detection for Danish, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 3498–3508. URL: https://aclanthology.org/2020.lrec-1.430/.

[11] E. Kaziaj, FUELLING hate: Hate speech towards women in online news websites in Albania, in: Gender and Sexuality in the European Media, Routledge, 2021, pp. 100–118.

[12] 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).

[13] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420. doi:10.18653/v1/N19-1144.

[14] M. S. Jahan, F. Hassan, W. Aransa, A. Bouchekif, Multilingual Hate Speech Detection Using Ensemble of Transformer Models, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[15] M. K. Sathya, K. Gopalakrishnan, M. PA, P. Balasundaram, Sinhala and gujarati hate speech detection, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[16] C. Muhammad Awais, J. Raj, Breaking Barriers: Multilingual Toxicity Analysis for Hate Speech and Offensive Language in Low-Resource Indo-Aryan Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[17] Y. Bestgen, Using Only Character Ngrams for Hate Speech and Offensive Content Identification in Five Low-Ressource Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[18] N. Narayan, M. Biswal, P. Goyal, A. Panigrahi, Hate Speech and Offensive Content Detection in Indo-Aryan Languages: A Battle of LSTM and Transformers, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[19] M. Rostamkhani, S. Eetemadi, Detecting hate speech and offensive content in english and indo-aryan texts, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[20] M. D. M. Qureshi, M. Sawant, M. A. Qureshi, W. Rashwan, A. Younus, S. Caton, Hate speech classification for sinhalese and gujarati, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[21] S. G GNANA, A. Venkatesh, K. N, O. M, B. V. A, P. Balasundaram, Enhancing hate speech detection in sinhala and gujarati: Leveraging bert models and linguistic constraints, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[22] S. Chanda, A. Dhaka, S. Pal, Crossing borders: Multilingual hate speech detection, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[23] G. Kalita, E. Halder, C. Taparia, A. Vetagiri, D. P. Pakray, Examining Hate Speech Detection Across Multiple Indo-Aryan Languages in Tasks 1 & 4, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[24] S. Agustian, Z. Idhafi, A. F. Rihardi, Improving detection of hate speech, offensive language and profanity in short texts with svm classifier, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[25] O. E. Ojo, O. O. Adebanji, H. Calvo, A. Gelbukh, A. Feldman, G. SIDOROV, Hate and offensive content identification in indo-aryan languages using transformer-based models, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.
[26] A. Joshi, R. Joshi, Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.