Overview of the HASOC Subtrack at FIRE 2023: Identification of Conversational Hate-Speech

Hiren Madhu[1], Shrey Satapara[2], Pavan Pandya[3], Nisarg Shah[3], Thomas Mandl[5] and Sandip Modha[6]

[1]*Indian Institute of Science, Bangalore, India*
[2]*Indian Institute of Technology, Hyderabad, India*
[3]*Indiana Bloomington University, USA*
[5]*University of Hildesheim, Germany*
[6]*LDRP-ITR, Gandhinagar, India*

Abstract

Identifying hate speech based on context is a requirement for real-world content moderation systems. However, in research, the definition and use of context for hate speech recognition has seen a variety of approaches. The task "Identification of Conversational Hate Speech" 2023 has provided a further dataset for hate speech detection, including context. The data was collected from Twitter (called X since 2023) and includes tweets, comments, or responses to such tweets or comments. This paper reports on the dataset, experiments, and results. Six teams submitted results for the binary classification task, and the best submission reached an F1 measure of 0.8. For the second task, five submissions were submitted. We also present a baseline that uses unlabelled data to obtain its predictions.

Keywords

Hate Speech, NLP, Social Media, Language Resource, Deep Learning, Text Classification, Evaluation, Benchmark, Context

1. Introduction

Hate speech and offensive language, which include hurtful, insulting, or derogatory remarks exchanged between individuals, are commonly observed on social media platforms like Facebook, Twitter, and Reddit. The abundant presence of such content on these platforms fosters offline hate crimes and fuels disorderly actions against various communities or political groups driven by agendas such as racism, misogyny, anti-LGBTQI+, anti-Muslim, anti-government, and other extremist ideologies [1]. To combat these hate crimes, the European Union (EU) and other European nations have implemented laws that classify online hate speech as a criminal offense, leading to the conviction of many individuals involved in such online activities. In contrast, the United States (US) primarily focuses on addressing hate speech through non-legal

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

hirenmadhu16@gmail.com (H. Madhu); shreysatapara@gmail.com (S. Satapara); pavanpandya1311@gmail.com (P. Pandya); nisarg0606@gmail.com (N. Shah); mandl@uni-hildesheim.de (T. Mandl); sjmodha@gmail.com (S. Modha)
0000-0002-6701-6782 (H. Madhu); 0000-0001-6222-1288 (S. Satapara); 0000-0002-8398-9699 (T. Mandl); 0000-0003-2427-2433 (S. Modha)

CEUR Workshop Proceedings (CEUR-WS.org)

means to safeguard the principles of free speech. While freedom of speech is crucial, a recent study [2] reveals that Elon Musk's influence on Twitter and his alterations to content moderation policies have increased hate speech on the platform. Consequently, this has caused numerous environmentally-conscious users to become inactive on the platform, resulting in a decline in the quality of discourse. In scenarios like this, freedom of speech acts as a double-edged sword.

Due to this, open societies need to figure out how to maintain civil discourse without resorting to totalitarian control, unlike the new Digital Service Act (DSA). DSA operates based on a "delete first, think later" approach, which removes user-generated content excessively and undermines freedom of expression [3]. Content moderation aims to strike a balance between these objectives. Nonetheless, moderating content necessitates numerous human annotators' involvement, making scalability impractical. This situation has driven research efforts toward the advancement of automatic systems for identifying harmful online content. Text classification represents just one component essential for meeting legal and practical demands [4], although it is crucial.

Most existing research in the field of identifying hate speech or offensive content primarily concentrates on analyzing the text of individual posts. Frequently, offensive or hateful content can be concealed within a conversation thread and may not be immediately evident in isolated comments or responses. However, it is feasible to uncover such hate speech by examining the original content and the context in which it was posted. Additionally, social media content often spans multiple languages, including code-mixed languages like Hinglish[1]. This makes it crucial for social media platforms to detect and remove such content before it reaches a wider audience. In the two editions of identification of conversational hate speech in code-mixed languages (ICHCL) [5, 6], datasets that handle such conversational hate speech have been released. The first edition featured binary labels to distinguish between hateful and regular tweets, while the second edition introduced a multiclass task, further categorizing hateful content tweets into two subtypes: standalone and contextual hate speech.

In this paper, we provide an overview of the third edition of ICHCL, which is centered on promoting the development of semi-supervised algorithms for classifying hateful text. Detailed information regarding the task and dataset is elaborated upon in Section 3.

2. Related Work

Because a tweet is typically a part of a larger discourse and a conversation among certain people, it is frequently difficult to understand it on its own. So far, only few text classification experiments and datasets took context into account. Context has been modeled in different ways. An early approach used LDA and RNNs.

Recursive neural networks were used to capture context within sentences [7] but less for capturing relations between subsequent messages in social media. LSTMs were used in an approach by Gao and Huang[8]. The dataset is based on comments in discussion threads on news articles and it contains 1500 comments. The context is given by the content of the news articles [8].

[1] Hindi written in Latin script instead of Devanagari script

The shared task RumourEval reacts to the need to consider evolving conversations and news updates for rumors and check their veracity [9]. The organizers provided a dataset of misinformation posts and conversations about those posts. The best performing system [10] used word2vec combined with several other dimensions such as source content analysis, source account credibility, reply account credibility and stance of the source message among others.

One dataset was labeled twice by crowd workers. One group was provided context and the other one not [11]. The data is extracted from Wikipedia talk pages. Context was given by the parent message and the title of the discussion thread [11]. It needs to pointed out that the parent message might not be the last message preceding the message to the annotated.

A further dataset that was extended with context information for the concept of abusiveness [12]. This data was collected based on an existing dataset without contextual information. For all tweets, the text was used to search them and if they were found, the authors tried to extract the previous messages. For all tweets, for which this was successful, the preceding messages were downloaded as context. Such a process leads to a greatly varying context size is very between items. Around 45% of the hateful tweets had one preceding tweet as context and another 45% had between 2 and 5 preceding tweets. Applying this methodology, almost half of the tweets which were annotated as abusive were labelled as non-abusive once context was available [12].

In a study with 10.000 Youtube comments, the quality of annotations in regard to interrater agreement was measured. Context improved the metrics by less than 5% in absolute terms [13].

In a study with Reddit posts, 27000 posts were annotated [14]. Context was given by providing the entire thread to the annotaters. However, diverse uses of context for annotation were reported. Quite low interrater agreement was reported, however, experiments showed an overall trend for improvement for context modeling [14].

Another dataset based of 6800 Reddit posts incuding the context of one preceeding comment was created [15]. A crowd worker annotation process showed low agreement and low quality annotations were disregarded. Showing the previous post changes the judgment in over 30% of the items of the Hate class. The best classification results reach F1 scores of 0.7 [15].

Within HASOC, datasets were collected in two previous editions of ICHCL and experiments were carried out [16, 6]. The diversity of the approaches, data sources and contexts definitions shows that further experiments are required.

3. ICHCL Task Overview and Dataset

A conversational thread might contain hateful, offensive, or profane language. This kind of content may not be immediately noticeable within individual tweets, comments, or responses to such tweets or comments. Nevertheless, it is possible to detect such hate speech by examining the original content and the context in which it was posted. For two editions of ICHCL, we have been focusing on detecting such hateful content in conversations. This year, we introduce a variation of the last two editions. We provide details about this in the following section. In the subsequent section, we discuss the details of the ICHCL 2023 dataset.

3.1. Task Overview

Training supervised models for classifying code-mixed text presents substantial challenges due to the limited availability of labeled data and the associated high cost of annotating large datasets. However, employing semi-supervised learning methods can alleviate these challenges by leveraging unlabeled data to improve model accuracy and reduce the need for extensive labeled data.

As a result the ICHCL task was developed further. Participants received an unlabeled training dataset and a labeled test dataset containing around 1,000 code-mixed Hindi samples. A crucial requirement was that participants must utilize the new unlabeled data to make predictions on the test dataset.

The classification task was divided into two subtasks:

- Task 2a: This subtask focuses on the binary classification of conversational tweets with tree-structured data into:
  - (NOT) Non-Hate-Offensive - This tweet, comment, or reply does not contain any hate speech or offensive content.
  - (HOF) Hate and Offensive - This tweet, comment, or reply contains hate speech, offensive, or profane content either on its own or in support of hate expressed in the parent tweet.
- Task 2b: This subtask is centered on classifying conversational tweets with tree-structured data into specific forms of hate, as follows:
  - (SHOF) Standalone Hate - This tweet, comment, or reply contains hate speech, offensive, or profane content on its own.
  - (CHOF) Contextual Hate - Comment or reply supporting the hate, offense, and profanity expressed in its parent. This includes affirming the hate with positive sentiment and having apparent hate.
  - (NONE) Non-Hate - This tweet, comment, or reply does not contain any hate speech or offensive or profane content.

This edition addresses the scarcity of labeled data and reduces annotation costs by providing only unlabeled data to participants. Participants received an unlabeled training dataset and a labeled test dataset comprising approximately 1,000 code-mixed Hindi samples. Additionally, a crucial requirement in this edition of ICHCL was that participants needed to leverage the unlabeled data which is provided to make predictions on the test dataset. Furthermore, a link to their GitHub repository to demonstrate compliance with the requirement was mandatory. To ensure fairness and equal opportunities for all participants, we imposed a condition that restricts the use to transformers with fewer than 200M parameters, preventing groups with extensive computational resources from gaining an unfair advantage.

3.2. Dataset

This section will provide an overview of how we gathered the dataset and present its statistics. To ensure a fair and unbiased sample of tweets, we selected controversial news stories covering a wide range of topics. We specifically handpicked contentious stories that were highly likely to contain hateful, offensive, or profane comments. These stories were drawn from the following categories:

- Brahmin Controversy in JNU
- Corruption
- Hinduphobia
- Kali smoking controversy
- Karnataka Election
- Kerala stories
- Modi clean chit
- Nupur Sharma
- Pakistan World Cup loss
- Udaipur murder
- Udhav Thakre government
- Zubair arrest

The participants were encouraged to use ICHCL 2021 and 2022 datasets as labeled data. In Table 1, we present the dataset statistics of the 2021 and 2022 datasets. We also present the statistics for the 2023 test data and the unlabeled training data. Table 2 presents the inter-annotator agreement for each level.

	#Twitter Posts		#Comments on Posts			#Replies on comments		
	HOF	NONE	HOF	CHOF	NONE	HOF	CHOF	NONE
Train (2021) [5]	49	33	1820	-	1958	972	-	908
	SHOF	NONE	SHOF	CHOF	NONE	SHOF	CHOF	NONE
Train (2022) [6]	75	97	588	171	1166	973	717	1127
Test	1	5	141	68	523	112	79	69
Total	125	135	2549	239	3647	2057	736	2104
Unlabeled	26		3928			4571		

Table 1
Dataset statistics for the ICHCL dataset

4. Results

In this edition of ICHCL, an initiative was put in place to encourage young researchers to develop innovative solutions. We introduced a semi-supervised version of the task to broaden the approaches used. Unfortunately, no submissions utilized the semi-supervised methods, highlighting a lack of interest in this part of the task. However, we still present the results and approaches by the participants.

As we can see in Table 3, for Task 2A, which focuses on binary classification, FiRC-NLP secured the top position with their submission ”parfirst2_all_folds,” achieving an impressive F1 score of 0.8079. They were closely followed by IRLab@IITBHU, Chetona, and AiAlchemists,

Type	IAA after two annotation rounds	IAA after three annotation rounds
Main	0.800	1.000
Comment	0.85381	0.80276
Replies	0.93961	0.90243

Table 2
Inter Annotator Agreement (IAA) for Task 2

Rank	Team Name	Submission Name	F1	Precision	Recall
1	FiRC-NLP [17]	parfirst2_all_folds	0.80791	0.80844	0.80741
2	IRLab@IITBHU[18]	IRLab@IITBHU_Task2A_1	0.70079	0.70255	0.69949
3	Chetona [19]	chetona-2a-def2	0.61551	0.62525	0.61425
4	AiAlchemists[20]	task2_binary_test_pred_2	0.61466	0.63351	0.60820
5	MUCS_3 [21]	MUCs_run_2	0.43474	0.38456	0.500
6	HASOC	BASELINE	0.37429	0.29909	0.500

Table 3
ICHCL Task 2A results

Rank	Team Name	Submission Name	F1 macro	Precision	Recall
1	FiRC-NLP [17]	parfirst_top3_top7_task2b	0.65414	0.64334	0.67178
2	IRLab@IITBHU[18]	IRLab@IITBHU_Task_2B_1	0.56316	0.56872	0.56685
3	AiAlchemists[20]	task_multiclass_1	0.38243	0.39198	0.39212
4	HASOC	BASELINE (Multiclass)	0.24952	0.19939	0.33333
5	Chetona [19]	chetona_2b_def2	0.17263	0.20795	0.15883

Table 4
ICHCL Task 2B results

demonstrating competitive results in precision and recall. In Task 2B, a multi-class classification task, FiRC-NLP continued to lead with their submission ”parfirst_top3_top7_task2b,” achieving a significant F1 macro score of 0.6541 presented in Table 4. IRLab@IITBHU and AiAlchemists also demonstrated notable performance. However, it is worth noting that the baseline submissions by HASOC in both tasks ranked lower, underlining the competitiveness of the shared task.

5. Methodology

In this section, we first explain the baseline we provided to the participants, and then we discuss the methodology of the top two teams.

5.1. Baseline Model

In order to support a low threshold for the entry to the shared task, a baseline model was provided for participants. It included with a template for steps like importing data, preprocessing, featuring, and classification. The participating teams could make changes in the code and experiment with various settings.

This year, we use a semi-supervised baseline, specifically, we use pseudo-labeling. First, we fine-tune a `bert-base-multilingual` on the labeled part of the dataset (2021, 2022 datasets). We then predict labels for the unlabeled training set (2023 training data) and then again fine-tune the model with the entire dataset (2021, 2022 datasets, and 2023 dataset with predictions).

5.2. Participant approaches

In this section, we explain and summarise the most successful participant approaches:

- **FiRC-NLP:** The sytsem uses concatenation to incorporate context and fine-tune XLM-RoBERTa-large for binary classification. For the multiclass task, the team fist applies the same binary classifier to classify hate and non-hate, and then fine-tunes another LLM to classify hate into standalone or contextual hate [17].
- **IRLab@IITBHU:** The submission implements a contrastive loss function to fine-tune the vanilla mBERT model, which is then used to obtain features for each individual level. After this step, they pass the features through a two-layer LSTM model to incorporate the context together with features from sentence BERT.
- **Chetona:** The submission concatenates the different levels of the conversational thread given. In then applies IndicBERT to encode the text and classifies based on the training data [19].

6. Conclusion

We reported on experiments with conversational and contextual hate speech detection. The new ICHCL dataset was created with a higher interrater agreement. The use of unlabelled data was set a the challenge for the task 2023. However, participants did not use the in that way. Overall, the submissions reached a good level of performance with up to a 0.8 F1 score applying deep learning models.

In future evaluations, data augmentation by large language models might be valuable directions. First experiments report positive outcomes [22].

References

[1] I. Kamenova, A. Perliger, 16. Online Hate Crimes, Handbook on Crime and Technology (2023) 278. doi:10.4337/9781800886643.00026.

[2] C. H. Chang, N. R. Deshmukh, P. R. Armsworth, Y. J. Masuda, Environmental users abandoned Twitter after Musk takeover, Trends in Ecology & Evolution (2023). doi:10.1016/j.tree.2023.07.002.

[3] A. Turillazzi, M. Taddeo, L. Floridi, F. Casolari, The digital services act: an analysis of its ethical, legal, and social implications, Law, Innovation and Technology 15 (2023) 83–106. doi:10.1080/17579961.202.

[4] A. Arora, P. Nakov, M. Hardalov, S. M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, I. Augenstein, Detecting harmful content on online platforms: What platforms need vs. where research efforts go, ACM Computing Surveys (2023). doi:10.1145/3603399, just Accepted.

[5] S. Satapara, S. Modha, T. Mandl, H. Madhu, P. Majumder, Overview of the HASOC subtrack at FIRE 2021: Conversational hate speech detection in code-mixed language, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 20–31. URL: https://ceur-ws.org/Vol-3159/T1-2.pdf.

[6] 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, 2022, pp. 475–488. URL: https://ceur-ws.org/Vol-3395/T7-1.pdf.

[7] H. Park, S. Cho, J. Park, Word RNN as a baseline for sentence completion, in: 5th IEEE International Congress on Information Science and Technology, CiSt 2018, Marrakech, Morocco, October 21-27, 2018, IEEE, 2018, pp. 183–187. doi:10.1109/CIST.2018.8596572.

[8] L. Gao, R. Huang, Detecting online hate speech using context aware models, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 260–266. doi:10.26615/978-954-452-049-6_036.

[9] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 845–854. doi:10.18653/v1/S19-2147.

[10] Q. Li, Q. Zhang, L. Si, eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 855–859. doi:10.18653/v1/S19-2148.

[11] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity detection: Does context really matter?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 4296–4305. doi:10.18653/v1/2020.acl-main.396.

[12] S. Menini, A. P. Aprosio, S. Tonelli, Abuse is contextual, what about NLP? the role of context in abusive language annotation and detection, CoRR abs/2103.14916 (2021). URL: https://arxiv.org/abs/2103.14916. arXiv:2103.14916.

[13] N. Ljubešic, I. Mozetic, P. K. Novak, Quantifying the impact of context on the quality of manual hate speech annotation, Natural Language Engineering 1 (2022) 14. doi:10.1017/S1351324922000353.

[14] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing cad: the contextual abuse dataset, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2289–2303. doi:10.18653/v1/2021.naacl-main.182.

[15] X. Yu, E. Blanco, L. Hong, Hate speech and counter speech detection: Conversational context does matter, arXiv Preprint (2022). URL: https://arxiv.org/abs/2206.06423.

[16] S. Satapara, S. Modha, T. Mandl, H. Madhu, P. Majumder, Overview of the HASOC subtrack at FIRE 2021: Conversational hate speech detection in code-mixed language,

in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 20–31. URL: http://ceur-ws.org/Vol-3159/T1-2.pdf.

[17] M. S. Jahan, F. Hassan, W. Mohamed, A. M. Bouchekif, Multilingual hate speech detection using ensemble of transformer models, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India, December 15-18, 2023, CEUR-WS.org, 2023.

[18] S. Chandal, A. Dhaka, S. Pal, Crossing borders: Multilingual hate speech detection, 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-WS.org, 2023.

[19] N. Madani, S. Saha, M. Sullivan, R. Srihari, Hate Speech Detection in Low Resource Indo-Aryan Languages, 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-WS.org, 2023.

[20] 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.

[21] P. M, R. K, A. Hegde, K. G, S. Coelho, H. L. Shashirekha, Taming toxicity: Learning models for hate speech and offensive language detection in social media text, 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-WS.org, 2023.

[22] A. Anuchitanukul, J. Ive, L. Specia, Revisiting contextual toxicity detection in conversations, ACM J. Data Inf. Qual. 15 (2023) 6:1–6:22. URL: https://doi.org/10.1145/3561390. doi:10.1145/3561390.