Overview of the HASOC Subtrack at FIRE 2022: Identification of Conversational Hate-Speech in Hindi-English Code-Mixed and German Language

Sandip Modha[1], Thomas Mandl[2], Prasenjit Majumder[3], Shrey Satapara[4], Tithi Patel[1] and Hiren Madhu[5]

[1]LDRP-ITR, Gandhinagar, India
[2]University of Hildesheim, Germany
[3]DA-IICT, Gandhinagar, India
[4]Indian Institute of Technology, Hyderabad, India
[5]Indian Institute of Science, Bangalore, India

Abstract

This article provides an overview of a shared task to identify contextual hate speech in social media conversations. This task intends to analyze how context within a conversation in social media can be used to improve the recognition of hate speech and offensive language. Within the ICHCL task and data set, messages which seem normal when viewed in isolation might be interpreted as containing or supporting hate speech, profanity, or other forms of offensiveness depending on the surrounding context. ICHCL provides a testbed for experimenting methods for best using the context from the preceding messages. The second goal of ICHCL is to draw even more distinctions between standalone hatred and hate in its social and conversational context. The multi-class classification of such contextual postings was the focus of this subtask. Twitter was used to sample the data set. An annotation tool was specifically built to retrieve and annotate around 5,200 code-mixed postings in English, Hindi, and German. In task-1, 12 teams submitted a total of 41 experiments. In task-2, 25 contributions were submitted by 10 different groups. The Macro-F1 score is the main criterion for ranking. The top-performing teams have reported a Macro-F1 score of 0.71 and a subtask score of 0.49, respectively. The task demonstrates how taking context into account may boost classification results.

Keywords

Conversational Hate Speech, Social NLP, Social Media, Deep Learning, Evaluation, Context

1. Introduction

People across the globe widely use social media like Twitter and Facebook due to their ease of use and the potential to network with others. The freedom to express oneself is a key benefit of these media systems. However, due to the anonymity and the social distance in digital

Forum for Information Retrieval Evaluation, December 9-13, 2022, India
sjmodha@gmail.com (S. Modha); mandl@uni-hildesheim.de (T. Mandl); p_majumder@daiict.ac.in (P. Majumder); shreysatapara@gmail.com (S. Satapara); tithigptd2000@gmail.com (T. Patel); hirenmadhu16@gmail.com (H. Madhu)
0000-0003-2427-2433 (S. Modha); 0000-0002-8398-9699 (T. Mandl); 0000-0001-6222-1288 (S. Satapara); 0000-0002-6701-6782 (H. Madhu)

communication, Hate Speech and other toxic content are frequently seen on such platforms. [1]. The platforms typically impose few limits on user-generated content. Actors with an agenda to defame others' reputations may post false and insulting information about them. It is very important for these platforms to detect such hate material before it spreads and to remain accessible to a large audience. Regulatory frameworks need to account for the nuance between protecting free expression and stifling it. Consequently, a great need for algorithmic support for content moderation rose. Many systems have been developed to detect Hate Speech[2, 3, 4].

Most hate speech detection algorithms in research depend on the text of a post alone without taking into account any contextual information. The recognition is typically carried out in a binary task [5]. However, it is often not possible to decide whether a conversational thread contains hateful or offensive material from a single remark or a reply to a comment. Much rather that task is possible only when considering the content of the parent post. The content on social media platforms is disseminated in a huge number of languages, including code-mixed varieties like Hinglish. Consequently, research and systems are necessary in many languages. ICHCL is the first benchmark that established a contextual data set for research on contextual Hate Speech recognition. ICHCL can help determine the most effective strategies for achieving this objective.

2. Related Work

Many datasets for hate speech and toxic content identification have been proposed. Many of the data sets are available for English, however, recently shared tasks have created new data sets for various languages such as Kurdish [6], Ethiopian [7], Portuguese [8] and Slovak [9]. These data sets have influenced the creation of machine learning models to automatically detect offensive content, ranging from SVM models with traditional features to state-of-the-art transformer models.

As elaborated above, a standalone post can often be hardly interpreted because it is part of a larger discourse and part of a conversation between some users. Using additional context information from the conversation available or from the account is a realistic task for Hate Speech identification. However, only a few text classification experiments and datasets considered context for the class assignment. An early approach was recursive neural networks which were used to capture context within sentences [10] but less for capturing relations between subsequent messages in social media.

Some approaches use a late fusion of text features and some meta-features of the account to facilitate text classification tasks. For example, Wang [11] has implemented such a model for fake news detection. The last layers of a model concatenate information that was distilled by diverse systems and fed them into a classifier [11].

The SemEval conference and evaluation initiative introduced the shared task RumourEval in 2019 (Determining Rumour Veracity and Support for Rumours) [12]. RumourEval reacts to the need to consider evolving conversations and news updates for rumors and check their veracity. The best performing system in subtask B by [13] used word2vec [14] for word text featuring combined with several other dimensions such as source content analysis, source account credibility, reply account credibility, and stance of the source message among others.

The authors concatenated all of these features in one model and applied an ensemble approach for classification.

The notion of toxicity is sometimes used as a more general term than hate speech. An interesting study developed a dataset that was labeled with and without context by crowd workers [15]. Half of the messages were annotated observing only the text of the message and the other half was annotated with additional context. The percentage of toxic messages is low in this dataset and reaches a maximum of 6 percent. The performance in both sets is similar, however, this seems no convincing argument that context is not helpful for a classifier.

Another dataset that was extended with context information adopts the notion of abusiveness [16]. 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. Applying this methodology, almost half of the tweets which were annotated as abusive were labelled as non-abusive once context was available [16]. Xu and colleagues developed a model for checking whether code words are used in their common meaning or with different meaning that is intended to relate to a group [17]. Such code words which are known in a community and might be used to hide Hate Speech and avoid content moderation.

The ICHCL task was already offered at FIRE 2021 [18]. [19] benchmarked the ICHCL dataset using most of the text representation schemes and classifiers. The best- performing pipeline uses a fine-tuned SentBERT paired with an LSTM as a classifier. This pipeline achieves a macro F1 score of 0.892 on the ICHCL test dataset. Overall, 15 research teams participated in the shared task. The reported macro F_1 score ranges around 0.49 to 0.73. The best team, MIDAS [20] developed ensembles of three transformer models, namely IndicBERT, Multilingual-BERT, and XLM-RoBERTa and reported macro-F1 score around 0.729. The authors concatenated posts to represent the conversational dialogue. The next two teams, Super Mario [21] and IIIT Hyderabad [22] used models based on XLM-RoBERTa and reported a macro F1 score around 0.71 and 0.70 respectively. The majority of the teams used different variants of BERT such as multilingual BERT, and IndicBERT for the classification. Team PC1 adopted a completely different approach. The authors converted text in the Devanagari script to ASCII characters. The author claims that this will work for any language. These results [18] represents the state-of-the-art performance for contextual Hate Speech identification.

3. HASOC Task Overview and Dataset

People's support for the hateful, offensive or profane material in conversational threads on social media is not always obvious from a single tweet, remark, or reply to a comment, but may be unearthed by looking at the larger conversational thread or the parent tweet. The primary motivation for offering this task is to discover content that encourages the spread of toxic content on social media platforms.

The following subsections will discuss the task design and will present the data set.

3.1. Task Overview

In this section, we'll discuss the two tasks that were offered. They are as follows:

3.1.1. Task-1: ICHCL HINGLISH and GERMAN Codemix Binary Classification

This task focuses on identifying hate speech and offensive language offered in Hinglish and German. Participants are expected to categorize tweets into two classes: hateful and offensive (HOF) and non-hateful and offensive (NOT). The descriptions of the classes are as follows:

- **Non Hate-Offensive (NOT)**: This post does not contain any Hate speech, profane, offensive content.
- **Hate and Offensive (HOF)**: This tweet, comment, or reply contains Hate, offensive, and profane content in itself or supports hate expressed in the parent tweet.

This can be best described by Figure 1. The parent/source tweet shows health-related antipathy against the individual. The screenshots show three comments. Without the context of the parent tweet, the three statements would not be regarded as offensive (ie, they would be labelled NOT). However, if we consider the conversational context, we might conclude that the comments reinforce the abuse represented by the original tweet. Therefore, these remarks should also be considered as offensive (ie, they will be labelled HOF).

3.1.2. Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) - Multiclass Classification.

Despite the fact that task 1 opens us to new frontiers to be conquered, it has a few negatives. One of which is seen in Figure 2. All the levels in Figure 2 will be labelled as hate. However, as the text demonstrates, they are all standalone hate. None of them are supporting or contextual hate. The labelling method, however, prevents models from recognizing if a chat thread comprises just independent hatred or also contextual hatred. To make amends, we've introduced task 2. Task 2 involves distinguishing between standalone and contextual hate. The labels in task 2 are as follows:

- **Standalone Hate (SHOF)** - This tweet, comment, or reply contains Hate, offensive, and profane content in itself.
- **Contextual Hate (CHOF)** - Comment or reply is supporting the hate, offense, and profanity expressed in its parent. This includes affirming the hate with positive sentiment and having apparent hate.
- **Non-Hate (NONE)** - This tweet, comment, or reply does not contain Hate, offensive, and profane content in itself.

So now, the comments in Figure 1 will be labelled contextual hate (CHOF) and the main tweet will be labelled as standalone hate (SHOF). On the other hand, comment and reply from 2 will be labelled standalone hate (SHOF) including the main tweet. We hope that introduction of a new label will help the language models understand the difference between standalone and contextual hate.

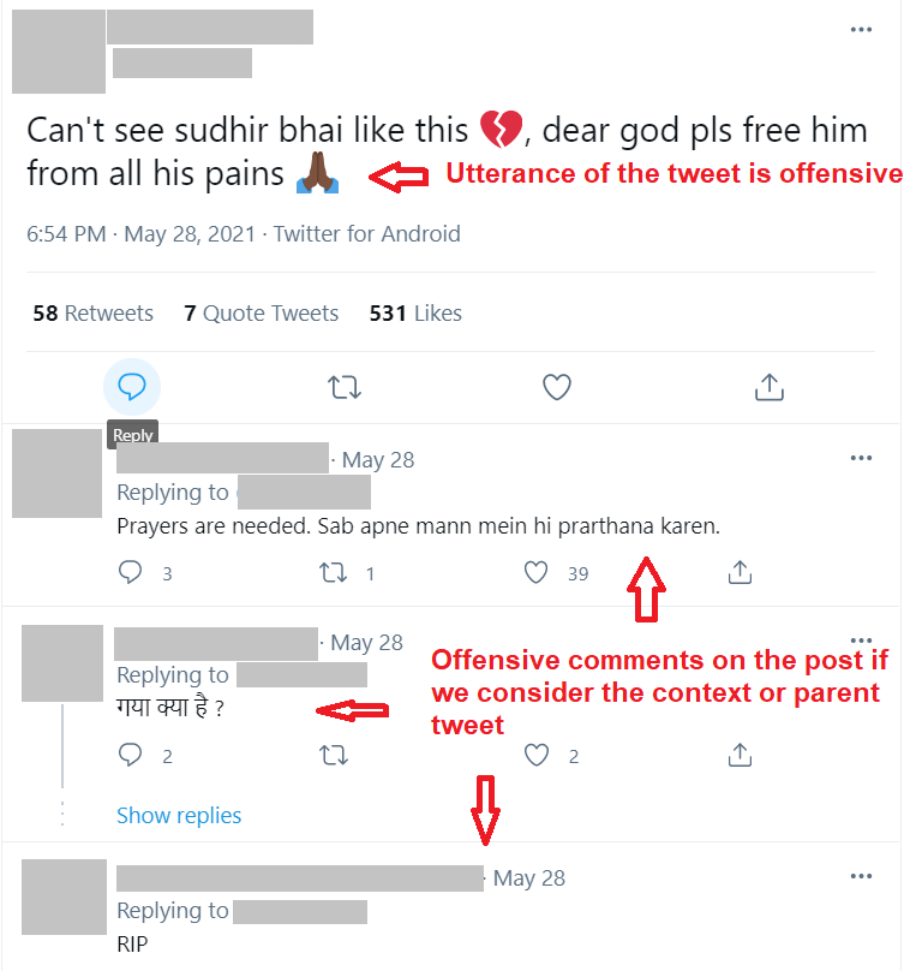

Figure 1: Conversational Hate Speech on Twitter

3.2. Dataset

In this subsection, we will present the dataset collection and dataset statistics. For sampling the tweets and to reduce the influence of prejudice, we have selected controversial stories on a variety of subjects. We've hand-selected controversial stories with a high likelihood of including hateful, offensive, and profane comments from the following categories. They are as follows:

- Bulli Bai App controversy
- Protest on Citizenship Amendment Act in India.
- Indian Celebrity Controversy.
- COVID-19
- Cast Controversy in India.
- Uniform Civil Code in India.
- Hinduphobia.

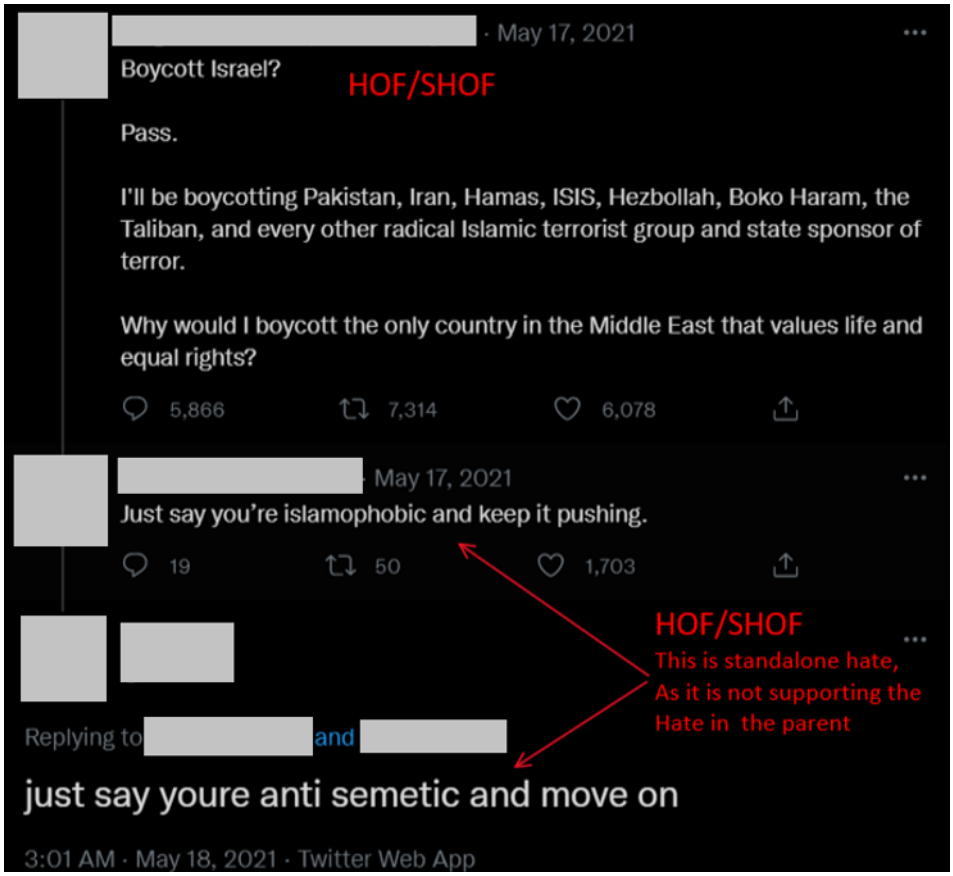

Figure 2: Misunderstandings in task 1

- Controversy related to Namaz in a public place.
- Farmer Protest on new farm laws in India.
- India -Pakistan Cricket match
- Differences in Indian history.
- Islamophobia.
- Russian-Ukrainian conflict.
- Antisemitism
- Kashmir issues.
- Ozil's Prayer on ground
- Temple-Mosque Controversies in India.
- Taliban takeover of Afghanistan

As a new task has been included, now the directory structure will also contain a contextual_labels.json file which will have labels for the task 2.

In terms of annotations, we only annotated task 2. This was because standalone hate and contextual hate together formed the HOF class in the binary classification dataset. We used a

Level	After 2 Annotations	After 3 Annotations
Main Tweets	0.672	1.0
Comments	0.6	0.95
Replies	0.667	0.91

Table 1
Level-wise Inter Annotator Agreement

different annotation algorithm for this task. In this algorithm, the annotations are done level by level. For example, first, all the main Tweets are annotated by two annotators, if there is a conflict then it is annotated by the third annotator. During this time, no comments or replies are annotated. Following this, all the comments are annotated by a max of four people and after all the comments are annotated without any conflict, the replies are annotated also by a max of four annotators. The inter-annotator agreement after two rounds of annotations was 0.51 and 0.95 for the remaining ones after 3 rounds of annotations. Table 1 presents the level-wise Inter annotator Agreement. After the third round of annotations, 95% of the tweets have no conflicts, indicating that the annotations are of high quality.

The tables 2 and 3 presents the dataset statistics of both the tasks.

Dataset		#Twitter Posts		#Comments on posts		#Replies on comments	
—		HOF	NOT	HOF	NOT	HOF	NOT
Train	Hinglish	75	97	759	1166	1690	1127
	German	6	5	59	136	23	78
Test	Hinglish	8	5	101	175	404	303
	German	2	2	20	40	1	16
Total		91	109	939	1517	2118	1524

Table 2
Dataset statistics for Task 1

	#Twitter Posts		#Comments on Posts			#Replies on comments		
	SHOF	NONE	SHOF	CHOF	NONE	SHOF	CHOF	NONE
Train	75	97	588	171	1166	973	717	1127
Test	8	5	76	25	175	266	138	303
Total	83	102	664	196	1341	1239	855	1430

Table 3
Dataset statistics for Task 2

4. Results

A total of 13 teams submitted 66 runs, 41 for task 1 and 25 for task 2. To give participants an idea of how to handle directory structure and contextual text, a baseline model was provided, which lowered the entry barrier. Details about the baseline model are described in section 5.1. Results of Baseline and teams are shown in Table 4 and 5. Figure 3 presents a comparison of the classwise F1 scores for CHOF, SHOF and NONE for task-2.

Table 4
Results of Task 1 Hindi-English Codemixed and German

Rank	Team Name	Macro F1
1	nlplab_isi [23]	0.7083
2	citk_isi [24]	0.6621
2	hate-busters [25]	0.661
3	bouchekif	0.6477
4	fosu-nlp [26]	0.6388
5	diu_bert	0.6281
6	irlab@iitbhu [27]	0.6271
7	sakshi hasoc [28]	0.6088
8	diu_bd_bert	0.6083
9	ml_ai_iiitranchi [29]	0.6008
10	uncle's boys	0.6004
11	HASOC (Baseline Results)	0.5937
12	gunjan [30]	0.5693
13	nitk_it	0.4173

Table 5
Results of Task 2 Hindi-English Codemixed

Rank	Team Name	Macro F1
1	ub-cs [31]	0.4939
2	HASOC (Baseline Results)	0.4899
3	fosu-nlp [26]	0.4769
4	bouchekif	0.4665
5	hate-busters [25]	0.4651
6	nlplab_isi [23]	0.4448
7	irlab@iitbhu [27]	0.439
8	ml_ai_iiitranchi [29]	0.4164
9	citk_isi [24]	0.3952
10	gunjan [30]	0.2865
11	sakshi [28]	0.2548

5. Methodology

In this section, we discuss the methodology used in the baseline model and the various approaches used by the participants.

5.1. Baseline Model

To reduce the barrier to entry into ICHCL and encourage participation from the scientific community, the organizers offered participants with a baseline model. The model implements TF-IDF, a traditional content representation method, and does not require any deep learning technologies. Participants could use and change this code, which includes feature design and

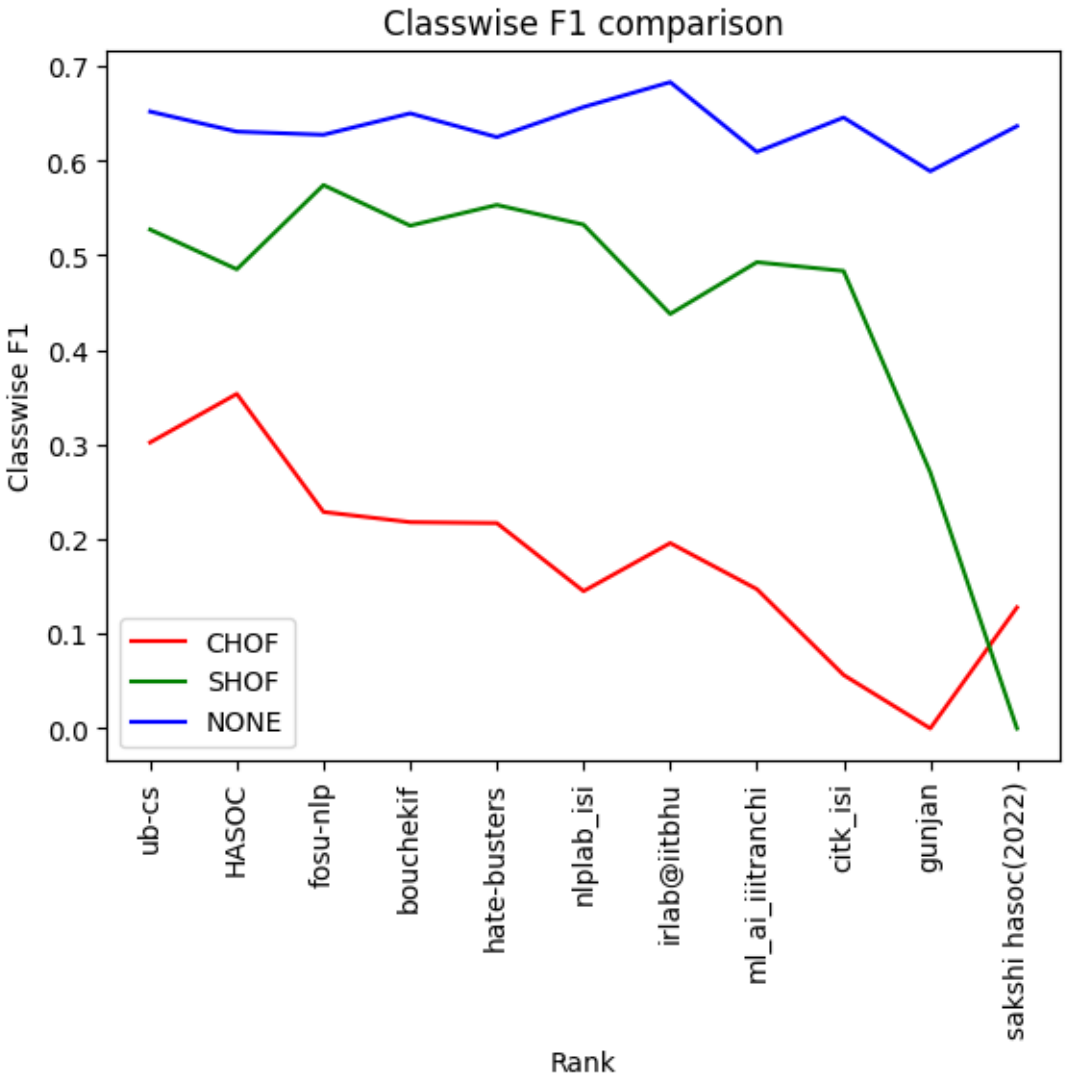

Figure 3: Classswise F1 scores

classification procedures, for their own studies. On a GitHub repository, the code for the basic model has been made accessible.[1].

The system architecture of the baseline model is as follows:

- First, all the libraries and the stemmer and stop words are loaded.
- All the JSON files are read.
- Then the concatenated tweets are vectorized using a TF-IDF Vecorizer from Scikit-Learn.
- These tweets are split into the validation set and the train set.

[1]https://github.com/hasocfire/ICHCL-baseline/tree/master/ICHCL_baseline-2k22

- A basic 2-layer MLP is trained on this dataset with 64 and 32 nodes in each layer respectively. These layers are followed by a classification module, with one or three nodes and sigmoid and softmax depending on the task. The model tries to minimize the binary cross-entropy or cross-entropy models depending on the task.
- This model is then trained on the dataset for 5 epochs with 32 batch sizes and Adam optimizer.

5.2. Participant systems

In this subsection, the system description of the top 5 teams from both tasks will be discussed.

5.2.1. Task-1

The top 5 teams in task 1 implemented the following systems:

- **nlplab_isi**: The levels were concatenated using '[SEP]' token. The system uses an ensemble of 3 fine-tuned transformer models, namely XLM-Roberta [32], Indic-BERT [33] and Google MuRIL [34]. However, using only one transformer was yielding better results compared to the ensemble model [23].
- **citk_isi**: The concatenated tweets are taken as input to a multilingual-Bert, which is fine-tuned by adding a classification layer at the end [24].
- **hate-busters**: This team uses an ensemble of three models. All used XLM-RoBERTa as the base but they are trained in three different ways. One of them is trained alone to optimize the isotropy property and to optimize the classification task as well. The other two models are ensemble models which rely on XLM-RoBERTa as the base but one of them uses 5-fold classification and another one is trained on the entire data using five different seeds. Similar to other approaches, the tweets were first concatenated [25].
- **fosu-nlp**: The tweets are concatenated in reverse order(ie, reply-comment-tweet) and 2 transformer models (one for Hinglish, one for German) were fine-tuned to achieve the classification [26].
- **irlab@iitbhu**: The concatenated tweets are taken as input to a XLM-RoBERTa, which is fine-tuned by adding a classification layer at the end [27].

5.2.2. Task-2

Out of the top 5 teams in task 2, only one is different from task 1 top 5 teams, and out of those 4 teams, only one has different systems for classification across tasks. Those are as follows:

- **ub-cs**: The best-submitted system "enhances" the tweets by augmenting the tweets to include emoji descriptions. This system first concatenates the levels and then fine-tunes an XLM-RoBERTa model. This with the "enhanced" tweets leads to the best F1 score for task 2 [31].
- **fosu-nlp**: The details of the implementation are identical to task-1, except that the team uses two different models for classification. The first model classifies between hate and non-hate while the second one classifies between standalone or contextual hate [26].

Most approaches simply integrate the context by concatenating the source tweet, the comment, and the reply. It might be a promising area of research to test further forms for considering the context.

6. Conclusions and Future Work

In comparison to ICHCL 2021, the performance stayed at the same level regarding the metric Macro-F1. The systems which were developed are rather similar. Mostly, transformer models were used. The baseline obtained the 2nd best performance for task 2. There might be two reasons for this observation. Many comments and replies of the category CHOF contained merely a few emojis.

The solution to concatenate the replies might lead to weak results due to their length. Furthermore, most teams remove emojis during preprocessing. Many short replies and comments will be transferred to an empty string. Thus, the preprocessing might not lead to beneficial information for the models. For the baseline, emojis were not removed, and using TF-IDF for features might give higher weights to emojis. This fact might explain the robust performance of the baseline. This observation is also supported by the fact that the best team in task 2 (ub-cs) translated emojis into a textual description. Thus, the semantic information within the emojis did not get lost. Furthermore, as we can see in figure 3, the F1 score for CHOF is lower compared to SHOF for all the teams. Also, the baseline model (rank 2) had the best F1 score (0.35) for the CHOF model with the rank 1 team (ub-cs) at a close second(0.30). Third team with F1 Score 0.22 has almost $\frac{1}{3}^{rd}$ of ranks above them. This further solidifies the fact that emojis play an important role in the detection of contextual hate.

The state of the art in research on contextual identification can only progress when further data sets are developed. The results of ICHCL 2022 show that the performance of classifiers for this task can still be improved.

Acknowledgments

We are thankful to Mr. Pavan Pandya and Mr. Jay Siddhpura for their contribution in developing the annotation and the HASOC-run submission platform. We are also thankful to all the annotators.

References

[1] R. Cohen-Almagor, Bullying, cyberbullying, and hate speech, Int. J. Technoethics 13 (2022) 1–17. URL: https://doi.org/10.4018/IJT.291552. doi:10.4018/IJT.291552.

[2] S. Modha, P. Majumder, T. Mandl, C. Mandalia, Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance, Expert Syst. Appl. 161 (2020) 113725. URL: https://doi.org/10.1016/j.eswa.2020.113725. doi:10.1016/j.eswa.2020.113725.

[3] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, Inf. 13 (2022) 273. URL: https://doi.org/10.3390/info13060273. doi:10.3390/info13060273.

[4] T. Chakraborty, S. Masud, Nipping in the bud: Detection, diffusion and mitigation of hate speech on social media, CoRR abs/2201.00961 (2022). URL: https://arxiv.org/abs/2201.00961. arXiv:2201.00961.

[5] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hatespeech and offensive content identification in English and Indo-Aryan languages, 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. 1–19. URL: http://ceur-ws.org/Vol-3159/T1-1.pdf.

[6] A. M. Saeed, A. N. Ismael, D. L. Rasul, R. S. Majeed, T. A. Rashid, Hate speech detection in social media for the kurdish language, in: K. Daimi, A. Alsadoon (Eds.), Proceedings of the ICR'22 International Conference on Innovations in Computing Research, Athens, Greece, 29-31 August, 2022, volume 1431 of *Advances in Intelligent Systems and Computing*, Springer, 2022, pp. 253–260. URL: https://doi.org/10.1007/978-3-031-14054-9_24. doi:10.1007/978-3-031-14054-9_24.

[7] W. B. Demilie, A. O. Salau, Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches, J. Big Data 9 (2022) 66. URL: https://doi.org/10.1186/s40537-022-00619-x. doi:10.1186/s40537-022-00619-x.

[8] F. Silva, L. A. de Freitas, Brazilian Portuguese hate speech classification using BERTimbau, in: R. Barták, F. Keshtkar, M. Franklin (Eds.), Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2022, Hutchinson Island, Jensen Beach, Florida, USA, May 15-18, 2022, 2022. URL: https://doi.org/10.32473/flairs.v35i.130594. doi:10.32473/flairs.v35i.130594.

[9] Z. Sokolová, J. Stas, D. Hládek, An introduction to detection of hate speech and offensive language in Slovak, in: 12th International Conference on Advanced Computer Information Technologies, ACIT 2022, Ruzomberok, Slovakia, September 26-28, 2022, IEEE, 2022, pp. 497–501. URL: https://doi.org/10.1109/ACIT54803.2022.9913104. doi:10.1109/ACIT54803.2022.9913104.

[10] 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.

[11] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL Vancouver, Canada, July 30 - August 4, Association for Computational Linguistics, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.

[12] 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.

[13] 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 Lin-

guistics, Minneapolis, Minnesota, USA, 2019, pp. 855–859. doi:10.18653/v1/S19-2148.
[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
[15] 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.
[16] 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.
[17] D. Xu, S. Yuan, Y. Wang, A. U. Nwude, L. Zhang, A. Zajicek, X. Wu, Coded hate speech detection via contextual information, in: J. Gama, T. Li, Y. Yu, E. Chen, Y. Zheng, F. Teng (Eds.), Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16-19, 2022, Proceedings, Part I, volume 13280 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 93–105. URL: https://doi.org/10.1007/978-3-031-05933-9_8. doi:10.1007/978-3-031-05933-9_8.
[18] 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.
[19] H. Madhu, S. Satapara, S. Modha, T. Mandl, P. Majumder, Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments, Expert Systems with Applications (2022) 119342. URL: https://doi.org/10.1016/j.eswa.2022.119342.
[20] Z. M. Farooqi, S. Ghosh, R. R. Shah, Leveraging transformers for hate speech detection in conversational code-mixed tweets, 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. 63–74. URL: http://ceur-ws.org/Vol-3159/T1-6.pdf.
[21] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, M. Das, Exploring transformer based models to identify hate speech and offensive content in English and Indo-Aryan languages, 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. 32–43. URL: http://ceur-ws.org/Vol-3159/T1-3.pdf.
[22] A. Kadam, A. Goel, J. Jain, J. S. Kalra, M. Subramanian, M. Reddy, P. Kodali, T. H. Arjun, M. Shrivastava, P. Kumaraguru, Battling hateful content in Indic languages HASOC'21, 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. 161–172. URL: http:

//ceur-ws.org/Vol-3159/T1-16.pdf.
[23] 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.
[24] 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: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[25] 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.
[26] H. Chen, Z. Han, L. Kong, Z. Zhang, Z. Li, M. Guo, H. Qi, Mixture Models based on BERT for Hate Speech Detection, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[27] S. Chanda, S. Sheth, S. Pal, Coarse and Fine-Grained Conversational Hate Speech and Offensive Content Identification in Code-Mixed Languages using Fine-Tuned Multilingual Embedding, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[28] S. Kalra, S. Goel, K. Maheshwari, Y. Sharma, Applying TF-IDF and BERT-based Variants for Hate Speech Detection in Code-Mixed Languages, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[29] K. Kumari, J. P. Singh, Machine Learning Approach for Hate Speech and Offensive Content Identification in English and Indo-Aryan Code-Mixed Languages, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[30] G. Kumar, J. P. Singh, HASOC Subtask at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[31] L.-D. Tebo, K. J. Ntwaagae, N. P. Motlogelwa, E. Thuma, M. Mudongo, Application of XLM-RoBERTa for Multi-Class Classification of Conversational Hate Speech, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2022.
[32] 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). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[33] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, P. Kumar, IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4948–4961. URL: https://aclanthology.org/2020.findings-emnlp.445. doi:10.18653/v1/2020.findings-emnlp.445.
[34] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, Muril: Multilingual representations for indian languages, 2021. arXiv:2103.10730.