Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German

Thomas Mandl
Information Science, University of Hildesheim
Hildesheim, Germany
mandl@uni-hildesheim.de

Sandip Modha
LDRP Institute of Technology and Research
Gandhinagar, India
sandip_ce@ldrp.ac.in

Anand Kumar M
Department of Information Technology, National Institute of Technology Karnataka
Surathkal, India
m_anandkumar@nitk.edu.in

Bharathi Raja Chakravarthi
Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland
Galway, Ireland
bharathi.raja@insight-centre.org

ABSTRACT

This paper presents the HASOC track and its two parts. HASOC is dedicated to evaluate technology for finding Offensive Language and Hate Speech. HASOC is creating test collections for languages with few resources and English for comparison. The first track within HASOC has continued work from 2019 and provided a testbed of Twitter posts for Hindi, German and English. The second track within HASOC has created test resources for Tamil and Malayalam in native and Latin script. Posts were extracted mainly from Youtube and Twitter. Both tracks have attracted much interest and over 40 research groups have participated as well as described their approaches in papers. In this overview, we present the tasks, the data and the main results.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Hate speech, datasets, evaluation, deep learning

ACM Reference Format:
Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. 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 (FIRE '20), December 16–20, 2020, Hyderabad, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441501.3441517

FIRE '20, December 16–20, 2020, Hyderabad, India
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8978-5/20/12…$15.00
https://doi.org/10.1145/3441501.3441517

1 INTRODUCTION

Hateful and offensive language use in online media is a social issue which has called much attention. The poisoning of online communication has been considered to be a serious problem for platforms. The classification of problematic posts has attracted much research. However, the problem is still challenging and there is a lack of research for many languages [18]. The track Hate Speech and Offensive Content Identification (HASOC) is dedicated to facilitate the development and improvement of classification systems for hate speech posts by providing data sets in several languages with few resources and to compare them to English. HASOC focused on Dravidian and Indo-European languages. The two sub tracks dealt with different language sets and are explained in the following sub sections.

2 HASOC INDO-EUROPEAN LANGUAGES

The first track within HASOC extends work from FIRE 2019 [13]. Further details are given in an exhaustive overview [14].

2.1 Tasks

The two primary tasks were offered again in 2020 while the third task was not offered again. The first task is a binary classification task. Systems need to find the problematic class (consisting of Hate Speech and Offensive posts).

- NOT: Non Hate-Offensive - This post does not contain any Hate speech, profane, offensive content.
- HOF: Hate and Offensive - This post contains Hate speech, offensive or profane content.

The second task is to further analyse the problematic class (HOF) into different types.

- HATE: Hate speech:- Posts under this class contain Hate speech content.
- OFFN: Offensive:- Posts under this class contain offensive content.
- PRFN: Profane:- These posts contain profane words.

Class	English	German	Hindi
NOT	1852	1700	2116
HOF	1856	673	847
PRFN	*1377*	*387*	*148*
HATE	*158*	*146*	*234*
OFFN	*321*	*140*	*465*
Sum	3708	2373	2963

Table 1: Statistical overview of the Training Data

Class	English	German	Hindi
NOT	391	392	466
HOF	423	134	197
PRFN	*293*	*88*	*27*
HATE	*25*	*24*	*56*
OFFN	*82*	*36*	*87*
Sum	814	526	663

Table 2: Statistical overview of the Development Data

2.2 Data

The data collection process has been modified to resemble the process at a platform. A large archive of tweets was processed and a preliminary classifier trained on hate speech collections is pre-selecting potentially problematic posts. These are then judged and labelled by humans. This process avoids a bias by searching a preliminary tweet set with hand-crafted keywords. The organisers have downloaded a large Twitter archive from the Internet Archive and extracted tweets in the languages which were relevant for these tasks. To obtain a set of tweets which contain potentially hateful tweets, we have trained SVM classifiers on the respective HASOC 2019 dataset and another additional dataset available for that language. For English, the OLID dataset was added [25]. For Hindi, the TRAC dataset was used [11]. For German, the GermEval 2018 dataset was added [24]. The purpose was to create a classifier that obtains an F1-score of around 0.5. These classifiers were applied to the dataset and all tweets classified as Hate speech went into the pool of documents for human judgment. In addition, 5% random tweets from the tweets not classified as Hate speech were also added to decrease bias.

Subsequently, assessment was carried out by students who were native speakers or typical users of social media in the respective language. All tweets were judged by at least two annotators. In cases of conflicts, when only 2 or 4 judgments were available and they disagreed, an algorithm considering the overall agreement determined the annotator with the higher reliability and took that judgment. Disagreement was fond in between 16% and 34% of the cases in task 1. Some of the automatic decisions were checked by a third annotator and she found the majority of decisions to be adequate. The amount of the final data is shown in Table 1 and Table 2. The development data represents 15% and another 15% were used for testing.

The submission and evaluation of experiments were handled on Codalab [1].

Rank	Team Name	Entries	F1 Macro average
1	NSIT_ML_Geeks	1	0.5337
2	Siva	1	0.5335
3	DLRG	2	0.5325
4	NITP-AI-NLP	1	0.5300
5	YUN111	1	0.5216
6	YNU_OXZ	2	0.5200

Table 3: Results of Top-Submissions for Subtask A - Hindi

Rank	Team Name	Entries	F1 Macro average
1	ComMA	4	0.5235
2	simon	1	0.5225
3	CONCORDIA_CIT_TEAM	1	0.5200
4	YNU_OXZ	3	0.5177
5	Siva	1	0.5158
6	Buddi_avengers	2	0.5121

Table 4: Results of Top-Submissions for Subtask A - German

Rank	Team Name	Entries	F1 Macro average
1	IIIT_DWD	1	0.5152
2	CONCORDIA_CIT_TEAM	1	0.5078
3	AI_ML_NIT_Patna	1	0.5078
4	Oreo	6	0.5067
5	MUM	3	0.5046
6	Huiping Shi	6	0.5042

Table 5: Results of Top-Submissions for Subtask A - English

2.3 Results

The results overall prove that the task remains challenging. No F1 score above 0.55 could be achieved. These scores are lower than at HASOC 2019. The results for each language are provided in the Tables 3, 4 and 5.

The top teams are are close together. This shows that despite a variety of approaches that was used, no advantage of a particular technology was identified. Most participants used deep learning models and in particular transformer based architectures were popular. Variants of BERT like ALBERT were used much. The best systems for the tasks have applied the following methodology. The best submission for Hindi used a CNN with fastText embeddings as input [19]. The best performance for German was achieved using fine-tuned versions of BERT, DistilBERT and RoBERTa [10]. The best result for English is based on a LSTM which used GloVe embeddings as input [15]. Very heterogeneous approaches were the best for the single languages.

For Task B, the best systems reached 0.29 for German, 0.33 [22] for Hindi and 0.26 for English [9].

3 HASOC DRAVIDIAN LANGUAGES

The second track within HASOC is focused on Dravidian languages for which very few resources are available. Further details are provided in an extensive overview [6]. The goal of this task is to identify offensive language from a code-mixed dataset of comments/posts in

[1] https://competitions.codalab.org/competitions/26027

Dravidian Languages (Malayalam-English and Tamil-English) collected from social media. The comment/post may contain more than one sentence but the average sentence length within the corpora is 1. Each comment/post is annotated with offensive language label at the comment/post level. The task-1 dataset also shows class imbalance problems depicting real-world scenarios. The participants were provided with development, training and test dataset.

3.1 Tasks

3.1.1 Task1: This is a message-level label classification task. Given a YouTube comment in code-mixed Malayalam, systems have to classify it into offensive or not-offensive.

3.1.2 Task2: This is a message-level label classification task. Given a tweet or YouTube comments in Tanglish and Manglish (Tamil and Malayalam using written using Roman Characters), systems have to classify it into offensive or not-offensive.

3.2 Data

3.3 Data for Task 1

For Task 1, we downloaded data from YouTube comments. The comments were downloaded from movie trailers during 2019. All the comment from those movie trailers were downloaded using a YouTube comment scrapper [2]. We utilized these comments to make a dataset for offensive language identification classification dataset. The dataset contains all types of code-mixing such as mixing the scripts of Malayalam script and Latin script, mixing at the word level, mixing at inter-sentential and intra-sential [5, 7].

3.4 Data for Task 2

The Tamil code-mixed dataset for Task 2 was collected from the Twitter tweets and comments on the Helo App. We have considered only the comments/posts in the Latin characters. Malayalam dataset for Task 2 has collected from YouTube comments. The training dataset for the Tanglish and Manglish used for the Task 2 contained 4000 comments. In Tamil, 2997 comments were collected from Twitter and 1003 are from Helo App. Out of 4000 comments, 1980 comments are offensive and 2020 comments are not offensive. The test dataset consists of 940 comments on which 475 are offensive and 465 are not offensive. Malayalam training set contains 1953 offensive comments and 2047 not-offensive comments, whereas the test set consists of 512 offensive and 488 not-offensive comments. The comments were annotated manually and verified by experts. The comments were annotated with two tags - OFF (offensive comment) and NOT (not-offensive comment). The dataset was delivered to the participants in CSV format.

3.5 Results

In Malayalam Task 1, teams SivaSai@BITS and IIITG-ADBU shared the first position with an F1-score of 0.95. These two systems achieved precision and recall score of 0.95. Teams from CFILT-IITBOMBAY and SSNCSE-NLP achieved the second position with an F-score of 0.94. The top four teams attained F-score higher than 0.90. The difference between the evaluation scores of top teams is minuscule. Table 6 presents the results of the Malayalam Task 1. Teams placed in the first and second position utilized transformer-based model for classification of YouTube comments into OFF and NOT. Transliteration of Romanized text into the native script is also found to be effective as part of this method. Another important

TeamName	Precision	Recall	F-Score	Rank
SivaSai@BITS [22]	0.95	0.95	0.95	1
IIITG-ADBU [4]	0.95	0.95	0.95	1
CFILT-IITBOMBAY	0.94	0.94	0.94	2
SSNCSE-NLP [3]	0.94	0.94	0.94	2
CENMates [16]	0.93	0.93	0.93	3
NIT-AI-NLP [12]	0.93	0.93	0.93	3
YUN [8]	0.93	0.93	0.93	3
Zyy1510 [26]	0.93	0.93	0.93	3
Gauravarora [2]	0.92	0.91	0.91	4
WLV-RIT [20]	0.89	0.90	0.89	5
Kjdong(only not)	0.70	0.83	0.76	6
Ajees [1]	0.69	0.38	0.44	7

Table 6: Rank list based on F1-score with other evaluation metrics (Precision and Recall) for Malayalam Subtask 1

TeamName	Precision	Recall	F-Score	Rank
CENmates [16]	0.78	0.78	0.78	1
SivaSai [22]	0.79	0.75	0.77	2
KBCNMUJAL [17]	0.77	0.77	0.77	2
IIITG-ABDU [4]	0.77	0.76	0.76	3
SSNCSE-NLP [3]	0.78	0.74	0.75	4
Gauravarora [2]	0.76	0.72	0.74	5
CFILT [23]	0.74	0.70	0.72	6
NITP [12]	0.71	0.68	0.69	7
Ajees [1]	0.72	0.67	0.68	8
Baseline	0.69	0.68	0.68	8
YUN [8]	0.67	0.67	0.67	9
Zyy1510 [26]	0.68	0.67	0.67	9
CUSAT [21]	0.54	0.54	0.54	10

Table 7: Rank list based on F1-score with other evaluation metrics (Precision and Recall) for Malayalam Subtask 2

TeamName	Precision	Recall	F-Score	Rank
SivaSaiBITS [22]	0.90	0.90	0.90	1
SSNCSE-NLP [3]	0.88	0.88	0.88	2
Gauravarora [2]	0.88	0.88	0.88	2
KBCNMUJAL [17]	0.87	0.87	0.87	3
IIITG-ADBU [4]	0.87	0.87	0.87	3
Zyy1510 [26]	0.88	0.87	0.87	3
CENmates [16]	0.86	0.86	0.86	4
CFILT [23]	0.86	0.86	0.86	4
YUN [8]	0.85	0.85	0.85	5
NIT-AI-NLP [12]	0.84	0.84	0.84	6
Baseline	0.85	0.84	0.84	6
Ajees [1]	0.84	0.83	0.83	7

Table 8: Rank list based on F1-score with other evaluation metrics (Precision and Recall) for Tamil Subtask 2

[2] https://github.com/egbertbouman/youtube-comment-downloader

fact visible from the results is that Support Vector Machine classifiers with TF-IDF features also reach top positions. Other systems submitted to the task use deep learning models using Bidirectional LSTM, LSTM, CNN and ULMFiT.

In Malayalam Task 2, CENmates reached the first position with an F-score of 0.78. Teams SivaSai@BITS and KBCNMUJAL bagged the second place, and their F-score was 0.77. The scores of the top five teams are close. Team CENmates used TF-IDF features with character n-gram as features for classification using machine learning algorithms. SivaSai@BITS used the same approach followed for Malayalam Task 1 for this task also. Team KBCNMUJAL used character and word n-grams features for with machine learning classifiers. Other teams used transformer-based models and Deep Learning-based models. Table 7 presents the result of Malayalam Task 2.

In Tamil Task 2, team SivaSai@BITS is placed in first position with an F-score of 0.90. They used a transformer-based model for this task also. Team SSNCSE-NLP gained the second position. They used TF-IDF with character n-gram features for classification. Gauravarora, who also came in second position applied a pre-trained ULMFiT for the classification. Three teams reached third position with an F-score of 0.87. These three teams used entirely different features and classifiers for the prediction task. Team KBCNMUJAL uses character n-gram and word n-gram features for representing text with ensemble classifiers. Team IIITG-ADBU used the same model used for other tasks for this task also. Team Zyy1510 used an ensemble of BiLSTM, LSTM+Convolution and a Convolution for the classification of social media texts into OFF and NOT. Table 8 presents the results of Tamil Task 2.

When we analyse the models submitted to the Tasks, most of them used either transformer-based models or conventional machine learning classifier with TF-IDF features. The performance of the deep learning models such as Bidirectional LSTM, LSTM, and CNN was not up to the mark. Transformer-based models used BERT for generating the embeddings.

4 CONCLUSION

Both tracks within HASOC have shown the current state of technology for the identification of offensive language. Development needs to continue to improve algorithms for these tasks.

REFERENCES

[1] Ajees A P. 2020. Ajees@HASOC-Dravidian-CodeMix-FIRE2020. In *FIRE (Working Notes)*. CEUR.

[2] Gaurav Arora. 2020. Gauravarora@HASOC-Dravidian-CodeMix- FIRE2020: Pretraining ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection. In *FIRE (Working Notes)*. CEUR.

[3] Nitin Nikamath Balaji and B Bharathi. 2020. SSNCSE-NLP@HASOC-Dravidian-CodeMix- FIRE2020: Offensive Language Identification on Multilingual Code Mixing Text. In *FIRE (Working Notes)*. CEUR.

[4] Arup Baruah, Kaushik Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2020. IIITG-ADBU@HASOC-Dravidian-CodeMix-FIRE2020: Offensive Content Detection in Code-Mixed Dravidian Text. In *FIRE (Working Notes)*. CEUR.

[5] Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. European Language Resources association, Marseille, France, 177–184. https://www.aclweb.org/anthology/2020.sltu-1.25

[6] Bharathi Raja Chakravarthi, Anand Kumar M, John P. McCrae, B. Premjith, K.P. Soman, and Thomas Mandl. 2020. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *FIRE (Working Notes)*. CEUR.

[7] Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. European Language Resources association, Marseille, France, 202–210. https://www.aclweb.org/anthology/2020.sltu-1.28

[8] Kunjie Dong. 2020. YUN@HASOC-Dravidian-CodeMix-FIRE2020: A Multi-component Sentiment Analysis Model for Offensive Language Identification. In *FIRE (Working Notes)*. CEUR.

[9] Tochukwu Ezike and Manikandan Sivanesan. 2020. Chrestotes at HASOC 2020: Bert Fine-tuning for the Identification of Hate Speech and Offensive Language in Tweets. In *FIRE (Working Notes)*. CEUR.

[10] Ritesh Kumar, Bornini Lahiri, Atul Kr. Ojha, and Akanksha Bansal. 2020. ComMA@FIRE 2020: Exploring Multilingual Joint Training across different Classification Tasks. In *FIRE (Working Notes)*. CEUR.

[11] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1–11. https://www.aclweb.org/anthology/W18-4401

[12] Sunil Kumar, Abhinav Saumya, and Jyoti Prakash Singh. 2020. NITP-AINLP@HASOC-Dravidian-CodeMix-FIRE2020: A Machine Learning Approach to Identify Offensive Languages from Dravidian Code-Mixed Text. In *FIRE (Working Notes)*. CEUR.

[13] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. 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*. CEUR, 14–17. http://ceur-ws.org/Vol-2517/T3-1.pdf

[14] Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. In *FIRE (Working Notes)*. CEUR.

[15] Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2020. IIIT_DWD@HASOC 2020: Identifying offensive content in multitask Indo-European languages. In *FIRE (Working Notes)*. CEUR.

[16] Veena P V, Praveena Ramanan, and Remmiya Devi G. 2020. CENMates@HASOC-Dravidian-CodeMix-FIRE2020: Offensive Language Identification on Code-mixed Social Media Comments. In *FIRE (Working Notes)*. CEUR.

[17] Varsha Pathak, Manish Joshi, Prasad Joshi, Monica Mundada, and Tanmay Joshi. 2020. KBCNMUJAL@HASOC-Dravidian-CodeMix-FIRE2020: Using Machine Learning for Detection of Hate Speech and Offensive Codemix Social Media text. In *FIRE (Working Notes)*. CEUR.

[18] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* (2020), 1–47. https://doi.org/10.1007/s10579-020-09502-8

[19] Roushan Raj, Shivangi Srivastava, and Sunil Saumya. 2020. NSIT & IIITDWD @ HASOC 2020: Deep learning model for hate-speech Identification in Indo-European languages. In *FIRE (Working Notes)*. CEUR.

[20] Tharindu Ranasinghe and Marcos Zampieri. 2020. WLV-RIT @ HASOC 2020: Offensive Language Identification in Code-switched Texts. In *FIRE (Working Notes)*. CEUR.

[21] Sara Renjit. 2020. CUSAT-NLP@HASOC-Dravidian-CodeMix-FIRE2020: Identifying Offensive Language from Manglish Tweets. In *FIRE (Working Notes)*. CEUR.

[22] Siva Sai and Yashvardhan Sharma. 2020. Siva@HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual Offensive Speech Detection in Code-mixed and Romanized Text. In *FIRE (Working Notes)*. CEUR.

[23] Pankaj Singh and Pushpak Bhattacharyya. 2020. CFILT IIT Bombay@HASOC-Dravidian-CodeMix FIRE 2020: Assisting ensemble of transformers with random transliteration. In *FIRE (Working Notes)*. CEUR. http://ceur-ws.org/

[24] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. (2018). https://ids-pub.bsz-bw.de/files/8493/Wiegand_Siegel_Ruppenhofer_Overview_of_the_GermEval_2018.pdf

[25] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. 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, 1415–1420. https://doi.org/10.18653/v1/N19-1144

[26] Yueying Zhu and Xiaobing Zhou. 2020. Zyy1510@HASOC-Dravidian-CodeMix-FIRE2020: An Ensemble Model for Offensive Language Identification. In *FIRE (Working Notes)*. CEUR.