Overview of the HASOC Subtrack at FIRE 2021: Conversational Hate Speech Detection in Code-mixed language

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

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

Abstract

This paper presents an overview of the newly developed subtask offered at the Forum for Information Retrieval (FIRE'21) conference on detecting contextual hate in social media conversational dialogue. Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) is offered as subtask-2 of the HASOC-English and Indo-Aryan Languages subtrack under the HASOC main track. The objective of the ICHCL subtask is to filter posts that are normal on a standalone basis but might be judged as hate, profane and offensive posts if we consider the context. This subtask focused on the binary classification of such contextual posts. The dataset is sampled from Twitter. Around 7000 code-mixed posts in English and Hindi were downloaded and annotated with an annotation platform developed for this task. A total of 15 teams from across the world has participated and submitted 50 runs for this track. The Macro F1 score is used as the primary metric for the evaluation. The best-performing team has reported a macro-f1 score of around 0.74. The task shows that considering the context can improve the performance of classification methods. ICHCL can contribute to identifying the best methods for this task.

Keywords
Hate Speech, NLP, context, social media

1. Introduction

Social media sites like Twitter and Facebook are free and highly user-friendly tools. They provide opportunities for people to air their voices. People, irrespective of age group, use these sites to share every moment of their lives which floods these sites with data. Apart from these positive features of social media, they have downsides as well. Due to the lack of restrictions set by these sites for their users to express their views as they like, anyone can make adverse and

unrealistic comments in abusive language against anybody with an ulterior motive to tarnish one's image and status in society. Regulation has to consider the thin line between free speech and censorship [1, 2, 3].

Most systems for hate speech detection are based on the text of a message merely without considering further elements [4, 5, 6]. A conversational thread can also contain hate and offensive content, which is not apparent just from a single comment or the reply to a comment but can be identified if the context of the parent content is provided. Furthermore, the content on such social media is spread in so many different languages, including code-mixed languages such as Hinglish [7]. So it becomes a huge responsibility for these sites to identify such hate content before it disseminates and remains visible for many users.

2. State of the Art: Collections for Contextual Analysis of Hate Speech

A single message in social media often cannot be interpreted alone because it appears as part of a larger discourse and part of a conversation between some users. Also for identifying Hate Speech, using the context could be beneficial. However, only a few text classification experiments and datasets considered context for the class assignment. There is a lack of collections that provide context for identifying messages better [8]. As a consequence, we provide a contextual task for the first time at HASOC 2021 [9] has along with plain vanilla classification task offered as tasks 1A and 1B [10] similar to the previous edition of HASOC [11, 12]

One approach to analyze tweets within a context and beyond their limited text content has been provided within SemEval 2019. The shared task RumourEval in 2019 (Determining Rumour Veracity and Support for Rumours) [13]. RumourEval reacts to the need to consider evolving conversations and news updates for rumors and check their veracity. The organizers provided a dataset of unreliable posts and conversations about those posts. Two tasks were offered. The second one (Subtask-B) was about verification of the rumour and it was modeled as a binary classification. The best performing system for this subtask used word2vec for representing text. It was combined with other knowledge about the tweet, the user, and the conversation. They considered source account credibility, reply account credibility, and stance of the source message among others. These features were concatenated in one model and entered into a classifier in parallel [14].

An approach closely related to hate speech detection is the detection of toxicity. The notion of toxicity can be used as a more general term than hate speech. A dataset from Wikipedia talk pages was labeled with and without context by crowd workers [15]. However, this collection still lacks a clear and stable definition of context.

Another dataset including context information for the notion of abusiveness was built based on an existing collection. 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 [16]. Almost half of the tweets which were annotated as abusive were labeled as non-abusive once context was available, which emphasizes the necessity for further research.

3. ICHCL Task Description

Social media users often support hate, offensive, and profane content in conversational threads, which is not visible in a single tweet, comment, or reply to a comment, but can be discovered if the context or parent tweet is considered. The main rationale behind offering this task is to identify such posts that support the dissemination of such impolite behaviour by the social media user.

Figure 1: Hate and Profane Conversational Dialogue Example

The screenshot from Twitter in the figure 1effectively describes the problem at hand. The parent/source tweet expresses abuse towards the individual regarding his personal health. Three

comments were seen in the screenshots. If the three comments were to be analyzed for the presence of hate or offensive speech without the context of the parent tweet, they would not be classified as offensive content. But if we take the context of the conversation into account, then we can say that the comments support the abuse expressed in the parent tweet. So those comments are labeled as offensive. The Figure 2is a screenshot of our annotation interface that describes the conversation hate speech problem.
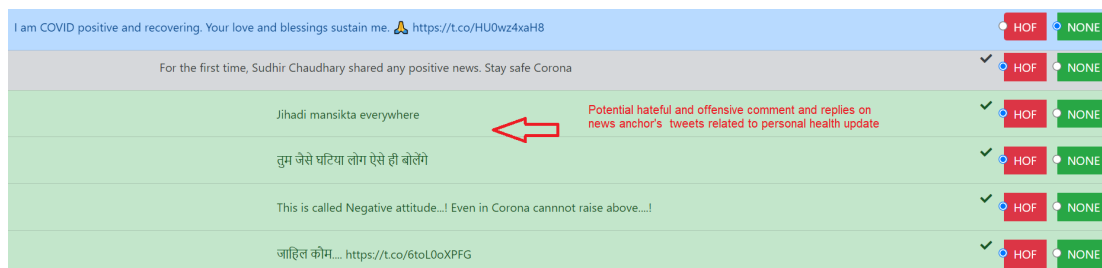

Figure 2: Hate and Profane Conversational Dialogue Example

This sub-task focused on the binary classification of such conversational tweets with tree-structured data into the following two classes:

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

4. The ICHCL Dataset

Sampling and annotating social media conversation threads is very challenging. A substantial amount of human intervention needed. In the following subsections, we describe the ICHCL dataset sampling, the annotation process and the dataset assembly.

4.1. Dataset Sampling

Using a scraper built with the Twitter API and the Selenium browser automation tool, we manually retrieved potentially problematic conversational chats from Twitter. We were able to scrape Twitter postings, comments on Twitter posts, and replies to each comment using this tool. We have chosen controversial stories on diverse topics to minimize the effect of bias. These were hand-picked controversial stories from the following topics that have a high probability of containing hate, offensive, and profane posts.

- Twitter Conflicts with the Indian Government on new IT rules
- Casteism controversy in India
- Charlie Hebdo posts on Hinduism

Class label	Training	Test
HOF	2841	695
NONE	2899	653
Total Tweets	5740	1348

Table 2
Training and Test Distribution of ICHCL Dataset

- The Covid-19 crisis in India 2021 Indian Politics
- The Israel-Palestine conflict in 2021
- Religious controversies in India
- The Wuhan virus controversy

Dataset	# Twitter Posts		#Comments the posts		#Replies	
	HOF	**NONE**	**HOF**	**NONE**	**HOF**	**NONE**
Train	49	33	1820	1958	972	908
Test	9	7	433	416	253	230
Total	58	40	2253	2374	1225	1138

Table 1: ICHCL Dataset: Training and Test set detail statistics

Table 1 presents the detailed statistics of the dataset. Table 2 displays the consolidated class distribution of the dataset.

4.2. ICHCL Dataset Annotations

Table 1 shows the amount of training and test data that was made available for this task. The conversation dialogues were extracted from Twitter using a targeted sampling approach. To ensure a high degree of quality, no crowd workers were used to annotate the dataset; instead, only the authors and a pre-final year student carried out the annotation. All tweets were annotated by at least two annotators. The conflict between the annotators is resolved by the third annotator. Labeling dialogue on social media is a pilot task and ICHCL was first introduced at FIRE 2021. For the annotation, we have developed our own software tool to annotate social media posts and dialogues. A sample video of the annotation system is available online[1]. Each dialogue, which includes posts, comments, and replies, is labeled as either HOF or NOT. The interrater agreement for the ICHCL task is around 74% and the Cohen coefficient is around 0.47.

---

[1]https://www.youtube.com/watch?v=DJq7OGdWRDE

4.3. Dataset Assembly

The Distribution of the tweets containing the conversation structure to the participants requires more data than for the HASOC subtask-1 [10]. Figure 3 shows the structure of the data directory:
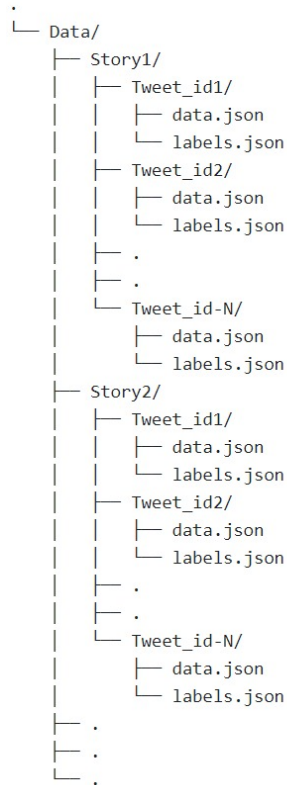
```
.
└── Data/
    ├── Story1/
    │   ├── Tweet_id1/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── Tweet_id2/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── .
    │   ├── .
    │   └── Tweet_id-N/
    │       ├── data.json
    │       └── labels.json
    ├── Story2/
    │   ├── Tweet_id1/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── Tweet_id2/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── .
    │   ├── .
    │   └── Tweet_id-N/
    │       ├── data.json
    │       └── labels.json
    ├── .
    ├── .
    └── .
```

Figure 3: Directory structure of the data directory

Figure 4 displays the structure of data.json file. The rectangles represent keys and ovals are elements containing arrays.

The content of the keys are as follows:

1. tweet: the text that is contained in the tweet
2. tweet_id: a global tweet_id generated by Twitter
3. comments: array of comments that a tweet has
4. replies: array of replies that a comment has

The structure of the labels.json file is flat. It contains no nested data structures. It only contains key-value pairs where the key is the tweet id and value is the label for the tweet with the given tweet id.
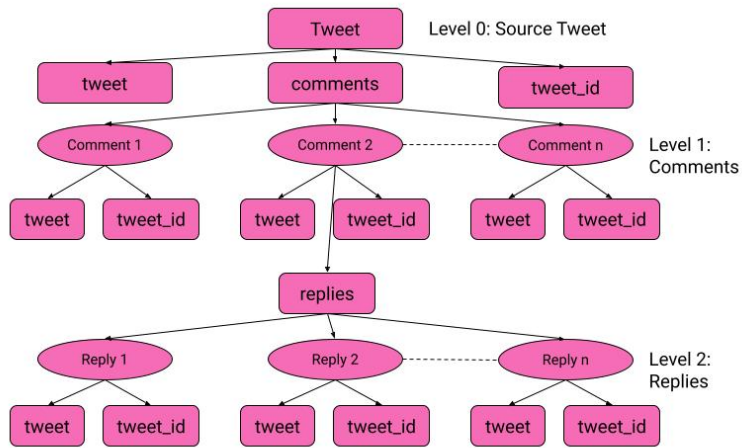
Figure 4: Structure of Data.json

5. Results

This subtask received a total of 50 submissions from 15 different teams. Furthermore, we provided the participants with a baseline model to give them a basic idea about the directory structure of the dataset and how to handle contextual text. Providing this code, the entry barrier was lower. The results of the teams and the baseline are shown in table 3.

6. Methodology

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

6.1. Baseline Model

To lower the entry barrier and motivate the community, the organizers decided to provide participants with a baseline model. Participants could use this code including feature design and classification processes and modify it for their experiments. The code for the baseline model is has been made open on a Github Repository[2].

The baseline system concatenates the comment after a source tweet and reply after a comment if any replies are present. Consequently, a comment has this structure: "<tweet> <comment>" and a reply looks the following one "<tweet> <comment> <reply>" in case the context is considered. For the rest of the paper, we will refer to this context representation as Concatenation.

---

[2]https://github.com/bhargav25dave1996/ICHCL_baseline

Rank	Team Name	# of runs	Macro F1
1	MIDAS-IIITD [17]	3	0.7253
2	Super Mario [18]	3	0.7107
3	PreCog IIIT Hyderabad [19]	3	0.7038
4	rider [20]	5	0.6890
5	Hasnuhana	4	0.6866
6	IRLab@IITBHU [21]	1	0.6795
7	r1_2021 [22]	5	0.6742
8	TeamBD [23]	5	0.6656
9	Chandigarh_Concordia	2	0.6551
10	PC1	5	0.6537
11	MUM [24]	5	0.6476
12	HASOC BASELINE	1	0.6315
13	TNLP [25]	4	0.6253
14	HUNLP [26]	2	0.6230
15	Oswald	1	0.5559
16	Sakshi HASOC [27]	2	0.4920

Table 3
Results of Subtask 2

Now after concatenation, a few simple preprocessing steps were applied like removing @handles, URLs, special symbols, etc. The text was vectorized using the TF-IDF weighting scheme and a simple Logistic Regression and Dense Neural Network were used as classifiers. This baseline model yielded a 0.6315 macro F1 score.

6.2. Methodologies

In this subsection, we briefly discuss the methodologies applied by the participating teams who have submitted track papers in the same order as they are presented in Table 3. The bullet points for the teams are created in the following form "Team_name <best_submission>".

- MIDAS-IITD <submit-3>[17]: Team MIDAS is the top team of ICHCL task. The authors proposed a transformer-based approach that relied on a concatenation of the contextual representation.They have used hard voting based ensembles of three transformer models, namely IndicBERT, Multilingual-BERT, and XML-ROBERta. The team added a dropout followed by a fully connected layer to the end each transfer-based model. Finally, the model combines probabilities of three models for the two classes, which were passed through a Softmax layer, and the scores were combined with an ensemble of classifiers using a hard voting scheme to obtain the final classification result.

- Super Mario <Context 1>[18]:The authors fine-tuned the XLM-Roberta-Large model with a classifier layer added at the end and trained on the ICHCL dataset. A binary cross-entropy scheme was applied for training the system.

- PreCog IIIT Hyderabad This team used XLM-RoBERTa for the classification.To capture the context and the tweet itself, the authors concatenate Twitter post comment, and replies using the [CLS] and [SEP] tokens. [CLS] and [SEP] are part of the vocabulary of

model, and are used to classify and take multiple sentences as input, respectively. The team used the CSNLI tool to convert the tokens in Latin script to Devanagari script

- rider <BERT-base-uncased>[20]: In this paper Authors approached are based on Multilingual Bert(Mbert), XLNet, Transfer of supervised features from a prominent English supervised dataset, and Ensemble Bert. Ensemble Bert is configured as the same hyperparameter as Bert with five random states. A classifier during the inference would take a vote of all five fine-tuned models in order to label a particular text sequence as a specific class. Authors reported best score using finetuned Bert-base-uncased model.

- r1_2021 <R1_v5> [22]:In this work, the authors produce experiments by using two context representation techniques. In the first method, child and parent tweets are merged and fed to the encoder, while in the second method, child and parent tweets are fed to separate encoders and output of encoder are averaged. Author have used Multilingual BERT and Indic-BERT. The best results were achieved by ensembling both BERT models and context representation techniques.

- TeamBD <LUT_DIU_Submission5>[23]: The authors expanded the ICHCL dataset using text augmentation methods. In particular, they applied automatic translation and back translations. The submitted experiments used BERT-base and CNN with TF-IDF weighted word embeddings. The latter induced better results for this team.

- PC1 <comboFeat>: This system uses a normalization process. The authors convert text in Devangari script to ASCII characters. The author claims that this will work for any language. TF-IDF was used to represent the character n-grams on the normalized text and classification was done using Logistic Regression.

- MUM <MUM_Task_2_4>[24]: The authors created representations for the text using several technologies. This included Emo2Vec, HastagVec, word uni-grams and char n-grams. Based on these schemes and features, they used an ensemble classifier with Random Forest, Gradient Boosting, MLP and with a soft voting to finally classify the tweet.

- TNLP <TNLP_CMH_S1>[25] : The authors experimented with an ensemble of several classifiers including Logistic Regression, Stochastic Gradient Descent, Naive Bayes, Random Forest, and Decision Tree classification models and used Indic BERT for obtaining the features.

7. Conclusion and Future Work

The task of conversational hate speech identification was introduced for the first time in this HASOC track. It received a reasonably good response from the NLP and AI community. We presented the dataset and a contextual baseline system based on a TF/IDF text representation and a SVM classifier [28]. Most of the teams outperformed the baseline results. Most of the approaches are based on different variants of BERT. Results show that considering the context increases the performance of Hate speech detection systems. However, it seems that there is more room for improvement and further elaborated methods for processing context.

In the next edition of HASOC, we intend to offer this task again and consider also other

languages. We would also like to consider multimodal features and summarizing [29]of conversational Twitter dialogues in the future.

8. Acknowledgments

References

[1] T. Gillespie, P. Aufderheide, E. Carmi, Y. Gerrard, R. Gorwa, A. Matamoros-Fernández, S. T. Roberts, A. Sinnreich, S. M. West, Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates, Internet Policy Review 9 (2020). doi:https://doi.org/10.14763/2020.4.1512.

[2] G. De Gregorio, Democratising online content moderation: A constitutional framework, Computer Law & Security Review 36 (2020) 105374. doi:https://doi.org/10.1016/j.clsr.2019.105374.

[3] S. Jaki, S. Steiger (Eds.), Digitale Hate Speech - Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation, Springer, Cham, 2022.

[4] S. Modha, T. Mandl, P. Majumder, D. Patel, Tracking hate in social media: Evaluation, challenges and approaches, SN Computer Science 1 (2020) 105. URL: https://doi.org/10.1007/s42979-020-0082-0.

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

[6] S. Modha, P. Majumder, T. Mandl, Filtering aggression from the multilingual social media feed, in: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), 2018, pp. 199–207.

[7] K. Bali, J. Sharma, M. Choudhury, Y. Vyas, ”i am borrowing ya mixing ?” an analysis of english-hindi code mixing in facebook, in: Proceedings of the First Workshop on Computational Approaches to Code Switching@EMNLP 2014, Doha, Qatar, October 25, 2014, Association for Computational Linguistics, 2014, pp. 116–126. URL: https://doi.org/10.3115/v1/W14-3914.

[8] V. Cotik, N. Debandi, F. M. Luque, P. Miguel, A. Moro, J. M. Pérez, P. Serrati, J. Zajac, D. Zayat, A study of hate speech in social media during the COVID-19 outbreak (2020). URL: https://openreview.net/forum?id=01eOESDhbSW.

[9] Modha, Sandip and Mandl, Thomas and Shahi, Gautam Kishore and Madhu, Hiren and Satapara, Shrey and Ranasinghe, Tharindu and Zampieri, Marcos, Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English

and Indo-Aryan Languages and Conversational Hate Speech, in: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[10] 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: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.

[11] T. Mandl, S. Modha, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of the Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE, CEUR-WS, 2019. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[12] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European Languages, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, volume 2826, CEUR-WS.org, 2020, pp. 87–111. URL: http://ceur-ws.org/Vol-2826/T2-1.pdf.

[13] 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, ACL, Minneapolis, Minnesota, USA, 2019, pp. 845–854. URL: https://www.aclweb.org/anthology/S19-2147.

[14] 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.

[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, Online, July 5-10., 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] F. Zaki, Mustafa, G. Sreyan, S. Rajiv, Ratn, Leveraging Transformers for Hate Speech Detection in Conversational Code-Mixed Tweets, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[18] 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: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[19] A. Kadam, A. Goel, J. Jain, J. S. Kalra, M. Subramanian, M. Reddy, P. Kodali, A. T. H, M. Shrivastava, P. Kumaraguru, Battling Hateful Content in Indic Languages HASOC '21, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[20] S. Mundra, N. Singh, N. Mittal, Fine-tune BERT to Classify Hate Speech in Hindi English Code-Mixed Text, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE),

CEUR-WS.org, 2021.

[21] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning Pre-Trained Transformer based model for Hate Speech and Offensive Content Identification in English, Indo-Aryan and Code-Mixed (English-Hindi) languages, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[22] R. Nayak, R. Joshi, Contextual Hate Speech Detection in Code Mixed Text using Transformer Based Approaches, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[23] M. S. Jahan, M. Oussalah, J. K. Mim, M. Islam, Offensive Language Identification Using Hindi-English Code-Mixed Tweets, and Code-Mixed Data Augmentation, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[24] A. Hegde, M. D. Anusha, H. L. Shashirekha, Ensemble Based Machine Learning Models for Hate Speech and Offensive Content Identification, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[25] R. Rajalakshmi, S. Srivarshan, F. Mattins, K. E, P. Seshadri, A. K. M, Conversational Hate-Speech detection in Code-Mixed Hindi-English Tweets, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[26] N. Bölücü, P. Canbay, Hate Speech and Offensive Content Identification with Graph Convolutional Networks, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[27] S. Kalra, K. N. Inani, Y. Sharma, G. S. Chauhan, Detection of Hate, Offensive and Profane Content from the Post of Twitter using Transformer-Based Models, in: Forum for Information Retrieval Evaluation (Working Notes) (FIRE), CEUR-WS.org, 2021.

[28] S. Modha, P. Majumder, T. Mandl, An empirical evaluation of text representation schemes to filter the social media stream, Journal of Experimental & Theoretical Artificial Intelligence (2021) 1–27. doi:https://doi.org/10.1080/0952813X.2021.1907792.

[29] S. Modha, P. Majumder, T. Mandl, R. Singla, Design and analysis of microblog-based summarization system, Social Network Analysis and Mining 11 (2021) 1–16. URL: https://doi.org/10.1007/s13278-021-00830-3. doi:10.1007/s13278-021-00830-3.

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