Overview of the HASOC Track at FIRE 2024: Hate-Speech Identification in English and Bengali

Nishat Raihan[1], Koyel Ghosh[2], Sandip Modha[3,4], Shrey Satapara[5], Tanishka Gaur[4], Yaashu Dave[4], Marcos Zampieri[1], Sylvia Jaki[6] and Thomas Mandl[6]

[1]George Mason University, Fairfax, VA, USA
[2]Indian Statistical Institute, Kolkata, India
[3]University of Milano-Bicocca, Milan, Italy
[4]LDRP-ITR, Gandhinagar, Gujarat, India
[5]Indian Institute of Technology, Hyderabad, India
[6]University of Hildesheim, Hildesheim, Germany

Abstract

Hate speech detection on social media platforms continues to be a major issue. It is challenging to detect hateful and offensive content due to a lack of datasets, particularly in languages with limited resources. To close this gap, benchmark datasets for these languages need to be developed. This research improves detection accuracy and offers information about how well offensive content is identified when compared to languages with more resources. To continue advancing research on low-resource languages, the Hate Speech and Offensive Content Identification (HASOC) shared task 2024 offered two tasks in Bengali and English. This paper outlines the objectives of the task, presents the datasets provided to participants, and presents an analysis of the participants' submissions. A total of 11 teams submitted runs to HASOC 2024. For English, the leading team achieved an F1 score of 0.813 and for Bengali the highest-performing team achieved an F1 score of 0.716. In HASOC 2024 a large variety of approaches were used by the participants including lexical approaches, transformer-based model as well as zero shot learning with LLMs.

Keywords

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

1. Introduction

Offensive speech is a common phenomenon in social media [1]. Detection and content moderation including deletion and down-ranking as measures are required to maintain a rational discourse for online users of platforms [2, 3]. The high prevalence of offensive and hate speech, for example, can be observed in the transparency database created by the EU which records deletion actions of platforms according to the Digital Service Act.[1]

Multiple survey and overview papers have been published on this topic in recent years evidencing the importance of creating system to recognize offensive content online [4, 5, 6, 7, 8, 9]. The initiative Hate Speech and Offensive Content Identification (HASOC) co-located with the Forum for Information Retrieval Evaluation (FIRE) has organized shared tasks on this topic since 2019 [10] creating important resources for several low resource languages [11].

HASOC 2024[2] focuses on identifying hate speech, offensive language, and profanity in Bengali and English. Bengali is a language spoken by over 230 million native speakers, mainly in the state West Bengal in India and in Bangladesh. The lack of resources for Bengali is also emphasized by Al Maruf et al.

Forum for Information Retrieval Evaluation, December 12-15, 2024, India
mraihan2@gmu.edu (N. Raihan); ghosh.koyel8@gmail.com (K. Ghosh); sjmodha@gmail.com (S. Modha); shreysatapara@gmail.com (S. Satapara); tangaur2507@gmail.com (T. Gaur); daveyaashu2411@gmail.com (Y. Dave); mzampier@gmu.edu (M. Zampieri); jakisy@uni-hildesheim.de (S. Jaki); mandl@uni-hildesheim.de (T. Mandl)
0000-0003-2427-2433 (N. Raihan); 0000-0003-2427-2433 (S. Modha); 0000-0001-6222-1288 (S. Satapara); 0000-0002-2346-3847 (M. Zampieri); 0000-0001-7840-7300 (S. Jaki); 0000-0002-8398-9699 (T. Mandl)

[1]The DSA Transparency Database can be found online: https://transparency.dsa.ec.europa.eu/
[2]https://hasocfire.github.io/hasoc/2024/index.html

[12]. The task involves classifying tweets into Hate and Offensive (HOF) or Non-Hate and Offensive (NOT). HASOC 2024 provides participants with TB-OLID, an existing Bengali dataset [13], and a new English dataset compiled for HASOC 2024. More details about the datasets are provided in Section 3.

The remainder of this paper is organized as follows. Section 2 presents an overview of related research. Section 3 describes the data and tasks included in HASOC 2024. Section 4 presents the results obtained by participants of the competition. Section 5 presents an analysis of the content of the datasets with the use of topic models. Finally, Section 6 concludes this paper and discusses avenues for future research.

2. Related Work

Many offensive language detection benchmarks are available for English and other high-resource languages. However, in the last few years, the NLP community has focused on creating more datasets for low-resource languages. The efforts for the creation of language resources for low-resource languages are of special importance. The aforementioned HASOC initiative has created resources for several languages of the Indian subcontinent. HASOC contributed datasets such as code-mixed Hindi [14], Gujarati [15], Tamil, Malayalam [16, 17, 18], Marathi [19], Assamese[20, 21], Bengali[20, 21], Bodo[20, 21], Gujarati[21], and Sinhala [21].

Hate speech detection quality depends on the datasets available for training. Potential biases needs to be identified in order to increase the generalization performance of the trained classifiers [22]. The framework introduced by Wich et al. [23] can be used to show the biases and characteristics of such datasets. This bias framework can quantify the difference of the probability distributions between and within hate speech datasets.

Bertram et al. [24] used several methods to analyze nine German hate speech datasets in order to gain insights into potential bias. Using different methods, the analysis shows the topical distribution of the different datasets. A recent study [25] analyzed six different English language hate speech datasets, with different but related labels like *hate speech*, *offensive*, *aggression* and *toxicity*. The authors visualized how similar and compatible classes are within and across the datasets and measured how well each class affects performance of hate speech classifiers. The results showed that even semantically similar classes varied and overlap with other related classes. They also imply that the performance of hate speech classifiers significantly depend on which class they were trained on. In annotation, even cultural background plays an important role [26].

Several other works explored hate speech datasets with regards to their biases and characteristics, as well as their generalizability. A study by Nejadgholi and Kiritchenko [27] explored two different types of bias in hate speech datasets and their effect on cross-dataset generalization: *topic bias* and *task formulation bias*. The former is a type of selection bias and was identified using keyword search. The authors showed that some topics are more generalizable than others. The latter bias describes the difference in the definitions of classes between the datasets. The effect of this bias was estimated by training classifiers on different tasks. The authors showed that in their setting, models tend to focus on specific terms and ignore the context.

A further important direction of research is the analysis of the performance of systems when one data set is used for training and others for testing [28]. Such results can also show how much the performance drops by using data from another distribution [29]. The drop also gives a hint on the capabilities to generalize the detection of hate speech.

3. Data and Task Description

3.1. English Dataset

We created a new dataset for English. The dataset was collected from X (Twitter). The language information provided by the platform was considered to filter out English tweets. The English task is

a coarse-grained binary classification in which participants were required to classify tweets into two classes, namely: hate and offensive (HOF) and non- hate and offensive (NOT) as described next:

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

The dataset contains 1,776 items. Some examples are shown in Table 1.

Table 1
Examples from the English dataset

Tweet	Offensive
@user @user Please urge our beloved President to skip the bill and just put us all back to work. We can handle the #chinavirus just fine. I'm a grandparent too and do not want an economic collapse over this virus. I'll take my chances so my children and grandkids can have jobs!	NOT
RT @user: so many girls think they're "bad bitch" like no you're just rude sit down	HOF
@user @user Very stupid comment made by an idiot.	HOF
Damn that was quick	NOT
Lot of staff in the office working from laptops. Get the fuck home.	HOF

3.2. Bengali Dataset

HASOC provided participants with TB-OLID [13], a Bengali dataset annotated following the Offensive Language Identification Dataset (OLID) taxonomy [30]. OLID considers whether an instance is offensive (level A), whether an offensive post is targeted or untargeted (level B), and what is the target of an offensive post (level C). As the second level of the TB-OLID annotation we consider OLID level A as follows:

- Offensive (O): Comments that contain any form of non-acceptable language or a targeted offense, including insults, threats, and posts containing profane language
- Non-offensive (N): Comments that do not contain any offensive language

Finally, the third level of the TB-OLID annotation merges OLIDs level B and C. We label whether a post is untargeted or, when targeted, whether it is labeled at an individual or a group as follows:

- Individual (I): Comments targeting any individual, such as mentioning a person with their name, unnamed persons or famous celebrities.
- Group (G): Comments targeting any group of people because of common characteristics, religion, gender, etc.
- Untargeted (U): Comments containing unacceptably strong language or profanities that are not targeted.

The statistics of the Bengali dataset is presented in Table 2. Overall, 1,000 Facebook posts/comments were labeled for the test set and 4,000 for the training set. An additional 500 instances are provided as a blind test for this shared task. Finally, some examples are shown in table 3.

4. Results

The results for the English task are presented in Table 4. A total of 21 systems were submitted by 8 teams. The best system reached an F1 score of 0.813. The following two systems are very close to the first system and both reached comparable performance.

Table 2
Statistical overview of the Bengali Data

Class	Bengali training	Bengali test	Bengali Blind-Test
Offensive	1954	427	237
Non Offensive	2046	573	263
Sum	4000	1000	500
Targeted at Group	806	148	115
Targeted at Individual	957	236	80
Untargeted	192	43	42
Code Mixed	1511	530	205

Table 3
Examples from the Bengali dataset

Tweet	Translated to English	Code Mixed Annotation	Offensive	Target
r k oek din por Bangladesh e o surgical strike chalabo.	After a long time, I will run a surgical strike in Bangladesh.	C	O	G
abaler dol sobkiso niye mithacar . tora moslim na hosh manosh hoo	Abal's party lies about all the casualties. You are not Muslim	T	O	G
Rubbish and stopid er moto kotha bola bad daoa uchit DADA.	Grandpa should be eliminated like Rubbish and Stopid.	C	O	I
Sala tui to akta janoar.tor vitor kono monusotto nei . bebek nai.	Shala you are a beast. You have no humanity inside. No conscience.	T	O	I
abar akbr prem a porlam re	Ray once again in love	T	N	
" Asbo kemne sob to gutibaaz ar dhanda baz.....ar jibon nosto korben "	"All these centers are all guts and dhanda lightning ..... and waste life".	T	O	U

Table 4
Results of the English task

Rank	Team	F1
1	FiRC-NLP	0.81333
2	Jing [31]	0.80050
3	wangkongqiang [32]	0.79385
4	TheNorthAtLTU [33]	0.76405
5	TextTitans [34]	0.75605
6	Abishikta	0.71038
7	MUCS [35]	0.5652
8	KK_IIITR_Research_Lab [36]	0.54924

For Bengali, 5 teams submitted 8 runs for task 1 and 7 runs for task 2. The best-performing system for the Bengali task 1 (offensiveness) reached a F1 score of 0.716. The following three teams obtained a similar performance level. The result for each team is displayed in table 5, ranked by their F1 scores.

The second task for Bengali was the classification of the target. There were fewer submissions for this second task. The results are given in table 6. For this task, much lower F1 values were achieved as it is more difficult.

The participants used a large variety of approaches. These start with classical methods as they were

Table 5
Results of Bengali (Level 1) task

Rank	Team	F1
1	FiRC-NLP	0.716
2	Falcon	0.7141
3	AIMLIIITRANCHI [37]	0.6845
4	MUCS [35]	0.6761
5	Initiators	0.3249

Table 6
Results of Bengali (Level 2) task

Rank	Team	F1
1	MUCS [35]	0.3975
2	FiRC-NLP	0.3218
3	Falcon	0.1871
4	AIMLIIITRANCHI [37]	0.1728

common before deep learning methods were established. Lexical features and supervised machine learning models were applied by Vinayak et al. [37]. Another supervised learning approach was adopted by Wang and Zhou [32] using tf/idf weighting and BERT embeddings. They used an external data set for training for the English task. The team also used augmentation and created additional tweets through deletion, shifting and substitution with synonyms.

Most teams utilized pre-trained transformer models to obtain embeddings. Supervised learning based on Universal Sentence Embeddings (USE) and LSTMs were applied by Alonso et al. [33]. Supervised learning with word features using classic supervised learning and boosting algorithms as well as deep learning (BiLSTM with max pooling) were applied by Kumari et al. [36]. Another group used BERT embeddings and fed them into a RNN and a CNN [31].

Diverse training sets were used for English. Alonso et al. [33] used three previous HASOC datasets and the OLID dataset. One team used a dataset from Kaggle and the HASOC 2021 dataset [36].

Team MUCS used BERT and DistilMBert used to obtain embeddings. The task description of HASOC was also embedded, and the cosine similarity between task definition and the tweets was calculated. This approach could be considered as a zero shot learning method because no training data was used [35].

The team TextTitans used GPT-3.5 Turbo for a zero shot approach. The authors used a simple prompt of two lines and changed the temperature setting to generate several runs. The performance differences were small [34].

Finally, one team also checked the relation between the predicted label and text length [33].

5. Dataset Analysis

For HASOC 2024, we also analyzed the content of the dataset to check whether any bias appears. We mapped the 4,000 tweets of the Bengali training set into a two-dimensional vector space in figure 1 using the TSNE model. The tweets were first translated automatically using the Google translate service from Bengali to English and then encoded with a SentenceTransformer using the 'stsb-distilbert-base' model. We can see that, at least in the TSNE model, the offensive and non-offensive items overlap considerably. The visual inspection implies that hate and non-hate posts do not simply fall into different thematic areas.

Furthermore, we provide a topic modeling analysis of the data sets. This allows a basic insight into the topics mentioned in the tweet collection.

Topic modeling is a technology for analyzing the content of a large collection of text documents [38]. For a human, topic modeling can lead to a good overview of content. The topics are presented as a

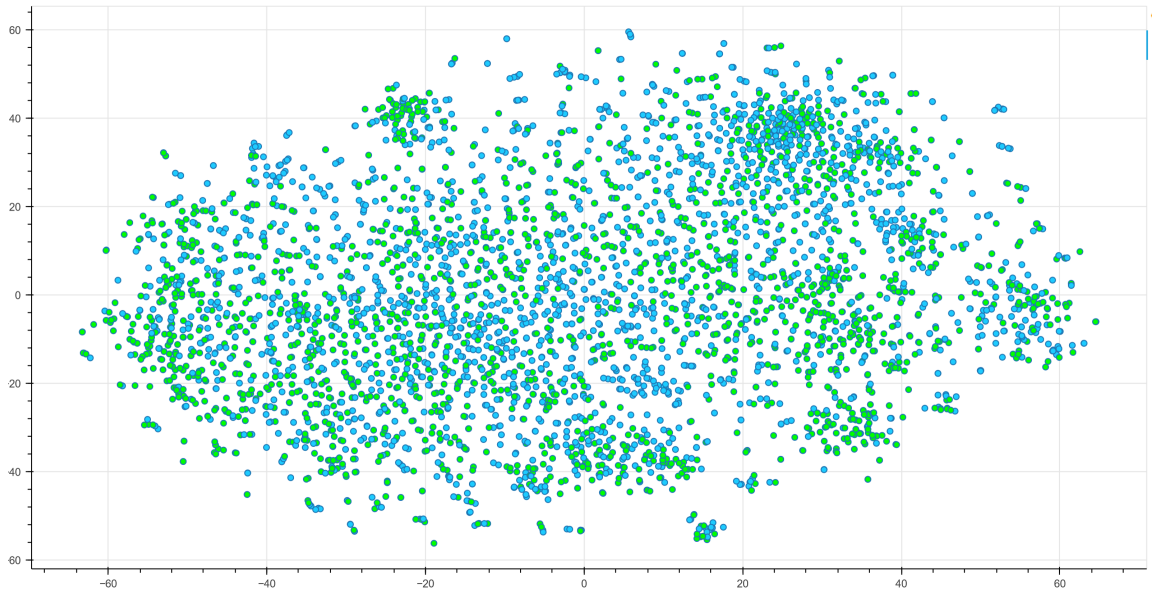

Figure 1: TSNE plot of the Bengali training set

collection of words which characterize this topic. Since topic modeling works unsupervised, it requires no training data, assumptions about content words and can be applied for exploring content without bias. BERTopic manages to maintain the semantic properties of documents better when compared to other approaches like LDA [39]. BERTopic provides a topic model that utilizes clustering techniques and weights based on term frequency and inverse document frequency (TF-IDF) values in order to obtain topics which take into account the semantic relationship between words [40].

BERTopic is based on the successful BERT transformer model [41] and utilizes its capacity for generating vector representations of words as well as sentences which represent the semantic content very well. BERTopic works by leveraging a pre-trained language model to create document embeddings which go through dimensionality reduction and clustering through Hierarchical Density-Based Spatial Clustering for applications with noise (HDBSCAN) [42]. The most relevant words of each cluster are classified through a class-based variation of TF-IDF [40].

We created topic models for the three datasets and heuristically searched for the most adequate number of topics. The large Bengali training set required the most topics, and the number was set to 15. The top-scoring topics are shown in figure 2. Only the first topics need to be reported as the frequency of documents is very high in the big topics, but drops drastically (see figure 3. The top topics of the test set are shown in figure 4. It can be observed that the topics do not overlap completely, but there are similarities. The major topics seem to be related to the relations between Bangladesh and India and other political issues.

The top scoring topics for the English dataset are shown in figure 5. It can be observed that this data contains tweets posted during the pandemic.

6. Conclusion and Future Work

We presented the results of the HASOC 2024 task on the detection of offensive language in Bengali and English. While this is the latest edition of HASOC, many open issues in hate speech research remain open. In multilingual countries, such as India, language resources still need to be developed to allow the development of systems capable of recognize offensive and hateful speech. The discussion on the quality of datasets needs to develop better measures for moving toward generalization. The detection of multi-modal content is also becoming increasingly relevant [43, 44].

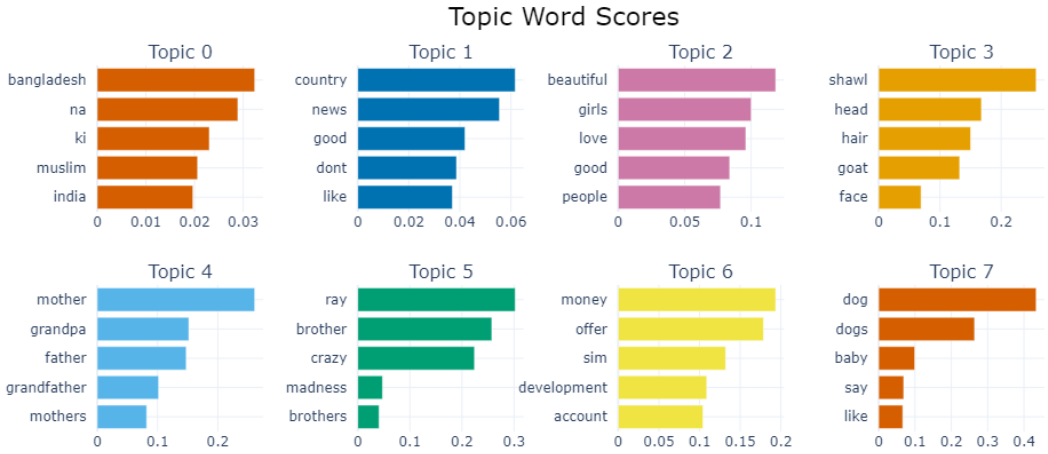

Figure 2: Top Topics of the Bengali training set

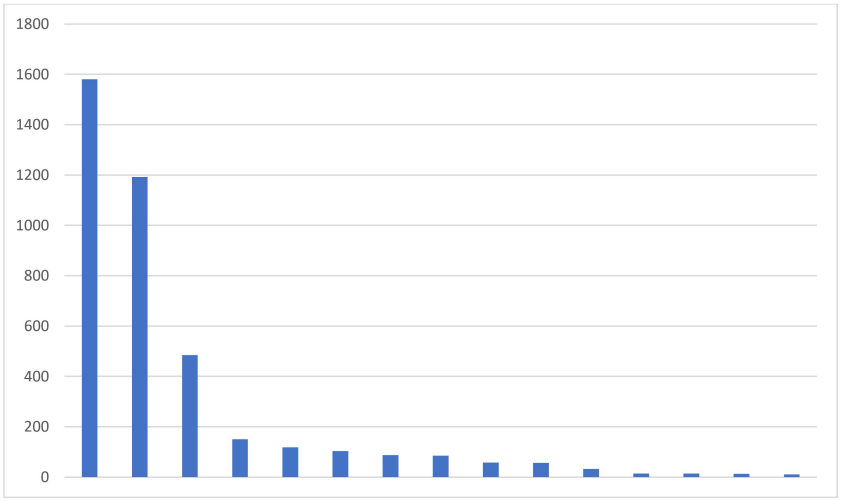

Figure 3: Size of the Topics in the Bengali training set

7. Acknowledgments

We would like to thank the annotators for their work. We further thank the shared task participants for submitting systems to HASOC 2024.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

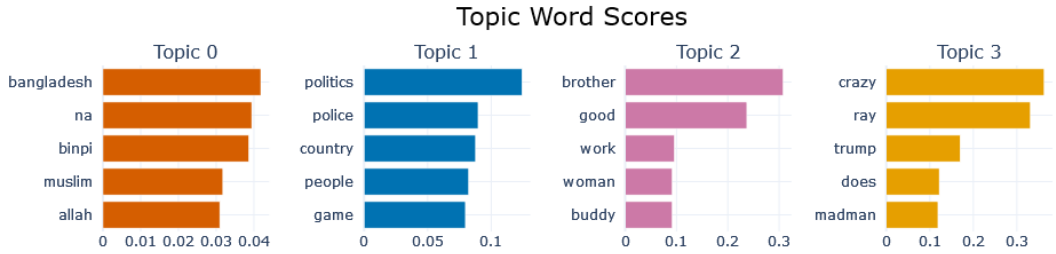

Figure 4: Top Topics of the Bengali test set

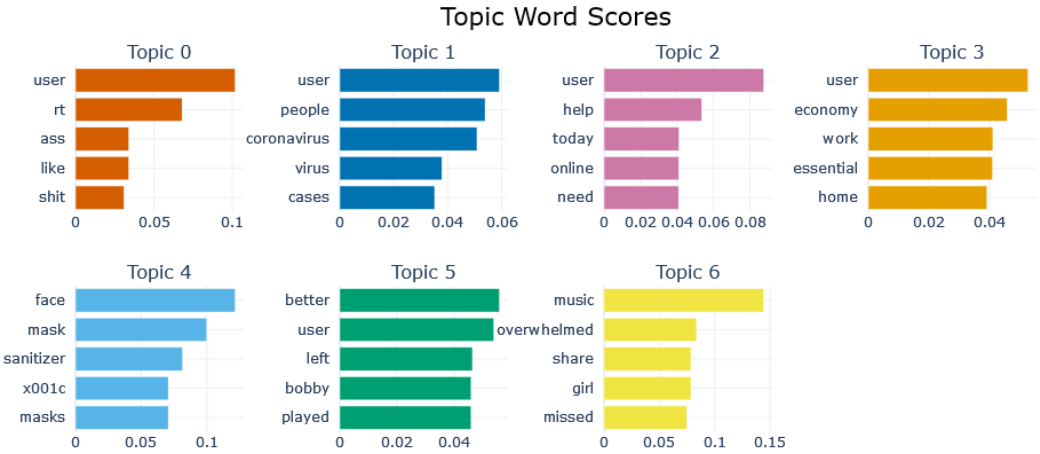

Figure 5: Top Topics of the English test set

References

[1] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, G. De Pauw, Online hatred of women in the incels.me forum: Linguistic analysis and automatic detection, Journal of Language Aggression and Conflict 7 (2019) 240–268. doi:10.1075/jlac.00026.jak.

[2] L. Cima, A. Trujillo, M. Avvenuti, S. Cresci, The great ban: Efficacy and unintended consequences of a massive deplatforming operation on reddit, in: Companion Publication of the 16th ACM Web Science Conference, Websci Companion '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 85–93. URL: https://doi.org/10.1145/3630744.3663608. doi:10.1145/3630744.3663608.

[3] T. Weerasooriya, S. Dutta, T. Ranasinghe, M. Zampieri, C. Homan, A. Khudabukhsh, Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 11648–11668.

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

[5] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, Neurocomputing 546 (2023) 126232. doi:https://doi.org/10.1016/j.neucom.2023.126232.

[6] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, A. Hussain, Hate speech detection: A comprehensive review of recent works, Expert Systems (2024) e13562. doi:10.1111/exsy.13562.

[7] A. Nandi, K. Sarkar, A. Mallick, A. De, A survey of hate speech detection in indian languages, Social Network Analysis and Mining 14 (2024) 70. doi:10.1007/S13278-024-01223-Y.

[8] M. Zampieri, S. Rosenthal, P. Nakov, A. Dmonte, T. Ranasinghe, Offenseval 2023: Offensive language identification in the age of large language models, Natural Language Engineering 29 (2023) 1416–1435.

[9] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, C. Silva, A comprehensive review on automatic hate speech detection in the age of the transformer, Social Network Analysis and Mining 14 (2024). doi:10.1007/s13278-024-01361-3.

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

[11] 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: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: https://ceur-ws.org/Vol-3159/T1-2.pdf.

[12] A. Al Maruf, A. J. Abidin, M. M. Haque, Z. M. Jiyad, A. Golder, R. Alubady, Z. Aung, Hate speech detection in the bengali language: a comprehensive survey, Journal of Big Data 11 (2024). doi:10.1186/s40537-024-00956-z.

[13] M. N. Raihan, U. Tanmoy, A. B. Islam, K. North, T. Ranasinghe, A. Anastasopoulos, M. Zampieri, Offensive language identification in transliterated and code-mixed Bangla, in: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), 2023, pp. 1–6.

[14] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, 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: D. Ganguly, S. Gangopadhyay, M. Mitra, P. Majumder (Eds.), FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021, ACM, 2021, pp. 1–3. doi:10.1145/3503162.3503176.

[15] S. Satapara, H. Madhu, T. Ranasinghe, A. E. Dmonte, M. Zampieri, P. Pandya, N. Shah, S. Modha, P. Majumder, T. Mandl, Overview of the HASOC subtrack at FIRE 2023: Hate-speech identification in sinhala and gujarati, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023, volume 3681 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 344–350. URL: https://ceur-ws.org/Vol-3681/T6-1.pdf.

[16] T. Mandl, S. Modha, A. K. M, B. R. Chakravarthi, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp. 29–32. doi:10.1145/3441501.3441517.

[17] K. Shanmugavadivel, M. Subramanian, P. K. Kumaresan, B. R. Chakravarthi, B. Bharathi, S. C. Navaneethakrishnan, L. S. Kumar, T. Mandl, R. Ponnusamy, V. Palanikumar, M. B. Jagadeeshan, Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, volume 3395 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 80–91. URL: https://ceur-ws.org/Vol-3395/T2-1.pdf.

[18] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, B. Premjith, S. K, S. C. Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, 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. 589–602. URL: https://ceur-ws.org/Vol-3159/T3-1.pdf. doi:10.5815/ijitcs.2021.03.03.

[19] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive language identification in marathi, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, volume 3395 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 489–501. URL: https://ceur-ws.org/Vol-3395/T7-2.pdf.

[20] K. Ghosh, A. Senapati, A. S. Pal, Annihilate hates (task 4 HASOC 2023): Hate speech detection in assamese bengali and bodo languages, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023, volume 3681 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 368–382. URL: https://ceur-ws.org/Vol-3681/T6-4.pdf.

[21] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in Assamese, Bengali, Bodo, Gujarati and Sinhala, in: D. Ganguly, S. Majumdar, B. Mitra, P. Gupta, S. Gangopadhyay, P. Majumder (Eds.), Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Panjim, India, December 15-18, 2023, ACM, 2023, pp. 13–15. doi:10.1145/3632754.3633278.

[22] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, Plos one 15 (2020) e0243300. doi:10.1371/journal.pone.0243300.

[23] M. Wich, T. Eder, H. A. Kuwatly, G. Groh, Bias and comparison framework for abusive language datasets, AI Ethics 2 (2022) 79–101. doi:10.1007/S43681-021-00081-0.

[24] M. Bertram, J. Schäfer, T. Mandl, Comparative survey of German hate speech datasets: Background, characteristics and biases, in: M. Leyer, J. Wichmann (Eds.), Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Marburg, Germany, October 9-11, 2023, volume 3630 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 207–221. URL: https://ceur-ws.org/Vol-3630/LWDA2023-paper19.pdf.

[25] P. Fortuna, J. Soler, L. Wanner, Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets, in: Twelfth Language Resources and Evaluation Conference, ELRA, Marseille, France, 2020, pp. 6786–6794. URL: https://aclanthology.org/2020.lrec-1.838.

[26] N. Lee, C. Jung, J. Myung, J. Jin, J. Camacho-Collados, J. Kim, A. Oh, Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2024. doi:10.18653/v1/2024.naacl-long.236.

[27] I. Nejadgholi, S. Kiritchenko, On cross-dataset generalization in automatic detection of online abuse, arXiv preprint arXiv:2010.07414 (2020).

[28] A. Dmonte, T. Arya, T. Ranasinghe, M. Zampieri, Towards generalized offensive language identification, arXiv preprint arXiv:2407.18738 (2024).

[29] P. Fortuna, Juan Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?, Information Processing & Management 58 (2021) 102524. doi:10.1016/j.ipm.2021.102524.

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

[31] J. Li, X. Yang, Hate Speech and Offensive Content Identification in English language based on BERT model, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation.

December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[32] K. Wang, X. Zhou, Two-step approach for Classification of Hate Speech and Offensive Content, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[33] P. Alonso, G. Kovács, R. Saini, M. Liwicki, Detection of Hate Speech using Universal Sentence Encoding and BiDirectional Long Short-Term Memory Models, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[34] A. Deroy, S. Maity, HateGPT: Unleashing GPT-3.5 Turbo to Combat Hate Speech on X, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[35] K. G, A. Hegde, S. D, Subrahmanya, H. L. Shashirekha, Zero-Shot and Multitask Learning Synergy for Robust Hate Speech Detection Across English and Bangla, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[36] K. Kumari, Avishikta, Vinayak, Hate Speech Detection for Hinglish Language, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[37] Vinayak, Avishikta, K. Kumari, U. K. Kedia, Hate Speech Detection for Bangla Language, in: Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation. December 12-15, Gandhinagar, India, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[38] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[39] R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology 7 (2022). doi:10.3389/fsoc.2022.886498.
[40] M. Grootendorst, BERTopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[41] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[42] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, volume 7819 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 160–172. doi:10.1007/978-3-642-37456-2_14.
[43] S. Jaki, T. Mandl, Memes in toxischer Online-Kommunikation. Ein Vergleich von genderbasierter Diskriminierung auf Tumblr und reddit, in: R. Opiłowski, H. E. H. Lenk, B. Mikołajczyk, N. Rentel (Eds.), Argumentation, Persuasion und Manipulation in Medientexten und -diskursen: 9. internationale Tagung zur kontrastiven Medienlinguistik: Argumentation, Persuasion und Manipulation in Medientexten und -diskursen. Universität Wrocław (Poland). 14.-16. September 2023, Vadenhoeck & Ruprecht unipress, Göttingen, 2024.
[44] M. Kalkenings, T. Mandl, University of Hildesheim at SemEval-2022 task 5: Combining deep text and image models for multimedia misogyny detection, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 718–723. URL: https://aclanthology.org/2022.semeval-1.98. doi:10.18653/v1/2022.semeval-1.98.