Baseline BERT models for Conversational Hate Speech Detection in Code-mixed tweets utilizing Data Augmentation and Offensive Language Identification in Marathi*

Koyel Ghosh[1], Apurbalal Senapati[1] and Utpal Garain[2]

[1]Central Institute of Technology, Kokrajhar, Assam, India
[2]Indian Statistical Institute, Kolkata, India

Abstract

In today's world, social media plays a vital role in spreading hate towards a person or group based on their color, caste, sex, sexual orientation, political differences, etc. Most of the work is done on a single tweet or comment classification, which lacks the conversation's context. The tweet, corresponding comments, and reply often helps us understand the context of the entire discussion. This paper discusses the used system and the performance of the team CITK_ISI on the first available code-mixed dataset on Hindi-English and German conversation scrapped from Twitter. Data augmentation is used with a baseline transfer-based BERT model and achieved a macro F1 score of 0.6653 for ICHCL Hinglish and German codemix binary classification. The system also identifies hate speech and offensive language in Marathi, a binary classification that secures a macro F1 score of 0.9019.

Keywords

Hate Speech, Transformers, Binary classification, Multiclass-classification, Code-Mixed Languages, Hindi-English, German, Marathi

1. Introduction

Instead of being friendly or informative, social media platforms like Twitter, Facebook, Youtube, etc. are becoming the platforms for cyberbullying and online harassment, leading people to depression or provoking people to involve in violence [1]. There are numerous instances around the globe in spreading such hate speeches disturbs social and communal integrity. As a result, numerous platforms of social media websites monitor user posts. This directs to an urgent injunction for methods to identify suspicious posts automatically. Most research on hate speech detection is done in English-like languages. Low-resource languages suffer from a lack of annotated datasets. Though few mono-lingual datasets in low-resource languages are available, code-mixed data like Hinglish (assembled of the words spoken in Hindi but written in the Roman script rather than the Devanagari script) are often used on Twitter, Facebook etc. This code-mixed language consists of different grammatical uses, slang and hateful words,

Forum for Information Retrieval Evaluation, December 9-13, 2022, India
*Corresponding author.
ghosh.koyel8@gmail.com (K. Ghosh); a.senapati@cit.ac.in (A. Senapati); utpal.garain@gmail.com (U. Garain)
https://github.com/BrainLearns (K. Ghosh)

including phonetic variations, misspelled words, and contextual usage in sentences. As well as the context of conversion plays a vital role in understanding the hate towards someone or something. Sometimes a parent tweet doesn't spread hate or fake news, but comments or replies associated with it directly attack the person who posts the tweet. Figure 1 shows an example reply supporting a hate comment towards a source tweet.

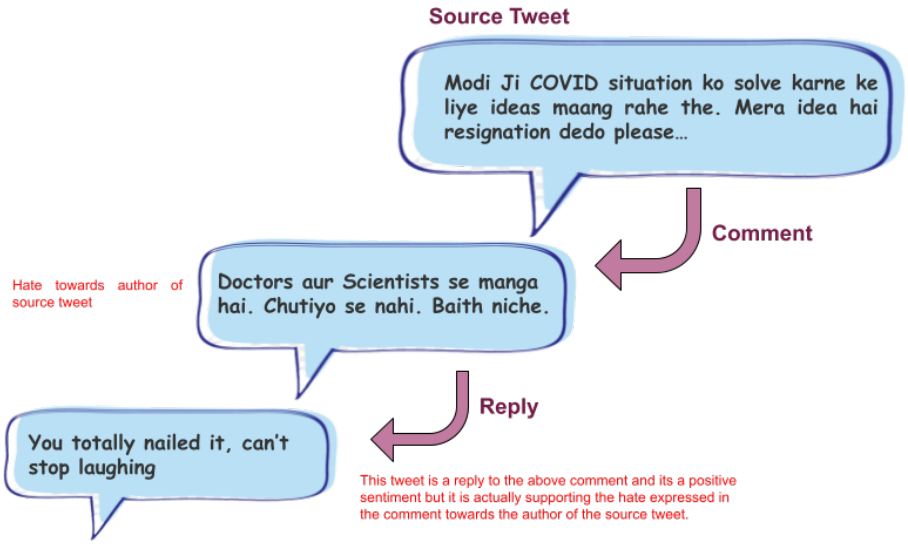

Figure 1: The reply has a positive sentiment "You totally nailed it, can't stop laughing.". But it is positive in favour of the hate expressed towards the author of the source tweet in the comment. Hence, it supports the hate expressed in the comment. Hence, it is also hate speech. The source tweet says, "Modi ji (PM of India) was asking for ideas to solve the covid situation of India. My idea to him is to resign." the comment expresses, "They have asked Doctors and Scientists. Not fuckers. Sit down."

Keeping this scenario in mind Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) 2022[1] [2] proposes two tasks this year 1) **Task 1 ICHCL Binary Classification** - Identification of Conversational Hate-Speech in Code-Mixed Languages like Hinglish and German 2) **Task 2 ICHCL Multiclass CLassification**- Identification of Conversational Hate-Speech in Code-Mixed Languages only in Hinglish. Along with that, they proposed 3) **Task 3A Marathi**- Offensive Language Identification in Marathi 4) **Task 3B Marathi**- Categorisation of Offensive Language in Marathi 5) **Task 3C Marathi**- Offense Target Identification in Marathi [3]. All five tasks are the extension of the previous year's HASOC 2021 task[2].

This paper attempted to identify hate speech content in all five tasks. Pre-trained BERT (Bidirectional Encoder Representations from Transformers) such as mBERT [4], MahaBERT [5], is used for this work.

The rest of the paper is structured as follows. Section 2 is the work related to hate speech detection in Hindi and Marathi languages. Section 3 describes the experimental setup, including the dataset, preprocessing steps, and baseline pre-trained BERT models. Section 4 shows the

[1] https://hasocfire.github.io/hasoc/2022/index.html
[2] https://hasocfire.github.io/hasoc/2021/index.html

results and findings from the experiments. Finally, it is concluded in Section 5.

2. Related work

The primary challenges of hate speech detection are the absence of related resources like language-specific datasets. Creating labeled datasets of hate speech in the Indian language is tedious and challenging. It needs lots of groundwork and preprocessing, like cleaning, annotators' agreements, etc., to create valuable data from social media. This section briefly outlines the existing approaches and available datasets of Hindi, Hindi-English and Marathi languages.

- **Hindi:** HASOC (Hate Speech and Offensive Content Identification), a shared task organized by FIRE (Forum for Information Retrieval Evaluation)[3], which published hate datasets in Indian languages such as Hindi, Marathi, etc. HASOC offers four subtracks, one of which is relevant to us: **HASOC - English and Indo-Aryan Languages**. The distribution of datasets comes in a tab-separated format. Other collections, including HASOC, use techniques to identify hate speech in online posts. In 2019, the HASOC-Hindi dataset offered three tasks [6]. Subtask A, which is the first task, is binary classification. Identifying the profanity or abuse (multiclass) of the hate comment is the second task, or subtask B. Subtask C involves determining if the hate speech is targeted at a specific person or is more general (untargeted). In Hindi, 93 runs were submitted for 3 different mini-tasks. Regarding the Hindi subtask A, the winner team, QutNocturnal [7], used a CNN-based method with a Word2vec embedding, yielding improved Marco F1 (0.8149) and Weighted F1 (0.8202) scores. The second group, LGI2P [8], employed BERT for classification after training a fastText model for the proposed Hindi language. Both the Marco-F1 and Weighted-F1 values for the system were 0.8111. Subtask B of the Hindi dataset receives a score of 0.5812 in Marco-F1 and 0.7147 in Weighted-F1 when BERT is used by 3Idiots [9]. This subtask C Hindi Dataset was completed with a high Marco-F1 score of 0.5754 by team A3-108 [10]. According to them, Adaboost [11] was the best performing classifier among the three classifiers, i.e., Adaboost or Adaptive Boosting (AB), Random Forest (RF), Linear Support Vector Machine (SVM). They merge multiple weak classifiers to construct a robust prediction model, but an ensemble of SVM, Random Forest, and Adaboost with hard voting performed even better. This classifier used TF-IDF features of word unigrams and characters 2, 3, 4, and 5 grams with an additional feature of the length of every tweet.
  In HASOC 2020, two Hate Speech detection tasks [12], sub-task A (binary class) and sub-task B (multiclass) are proposed with another Hindi dataset in the research area. NSIT_ML_Geeks [13] outperforms other teams in the competition scoring Marco-F1 0.5337 and 0.2667 in sub-task A and sub-task B, respectively, utilizing CNN and BiLSTM. Nohate [14] team achieved Marco-F1 0.3345 in sub-task B, fine-tuning the BERT model for the classification.
  In 2021, HASOC published a Hindi dataset [15] with sub-task A and B again. Total Sixty-five teams submitted a total of six thousand and fifty-two runs. The best submission

[3] http://fire.irsi.res.in/fire/2022/home

was achieved Macro F1 0.7825 in sub-task A with a fine-tuned Multilingual-BERT (20 epochs) with a classifier layer added at the final phase. The second team also fine-tuned Multilingual-BERT and scored Macro F1 0.7797. NeuralSpace [16] got Macro F1 0.5603 in sub-task B. They use an XLM-R transformer, vector representations for emojis using the system Emoji2Vec, and sentence embeddings for hashtags. After that, three resulting representations were concatenated before classification. In the paper [17] they used the pre-trained multilingual BERT (m-BERT) model for computing the input embedding on the Hostility Detection Dataset (Hindi) later SVM, Random-Forest, Multilayer perceptron (MLP), Logistic Regression models are used as classifiers. In coarse-grained evaluation, SVM reported the best weighted-F1 score of 84%, whereas they obtained 84%, 83%, and 80% weighted-F1 scores for LR, MLP, and RF. In fine-grained evaluation, SVM has the most excellent F1 score for evaluating three hostile dimensions, namely Hate (47%), Offensive (42%), and Defamation (43%). Logistic Regression beats the others in the Fake dimension with an F1 score of 68%.

- **Hindi-English:** In 2021, HASOC's main track had another subtrack, i.e., Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) [18], offered as subtask-2 of the HASOC-English and Indo-Aryan Languages subtrack. The ICHCL subtask aims 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. Team MIDAS [19] is the top team of the 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: IndicBERT, Multilingual-BERT, and XML-ROBERta. The team added a dropout followed by a fully connected layer to the end of each transfer-based model. Finally, the model combines the probabilities of three models for the two classes, passed through a Softmax layer. The scores were combined with an ensemble of classifiers using a hard voting scheme to obtain the final classification result. The authors of Super Mario [20] 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 to train the system.

- **Marathi:** In HASOC-Marathi [15], the best-performing team, WLV-RIT fine-tuned XLM-R Large model with a simple softmax layer. Later executed transfer learning from English data released for OffensEval 2019 [21] and Hindi data released for HASOC 2019 [6] and show that executing transfer learning from Hindi is better than executing transfer learning from English. They Scored an F1 score of 0.9144 [22]. The second team applied a fine-tuned LaBSE transformer [23] on the Marathi and the Hindi data set and achieved an F1 score of 0.8808. Their experiments show that the LaBSE transformer [24] outperforms XLM-R in the monolingual settings, but XLM-R performs better when Hindi and Marathi data are merged. L3CubeMahaHate [25] presents the first major Marathi hate speech dataset with 25,000 distinct tweets from Twitter, later annotated manually, and labeled them into four major classes, i.e., hate, offensive, profane, and not. Finally, they use CNN, LSTM, and

Transformers. Next, they explore monolingual and multilingual variants of BERT like MahaBERT, IndicBERT, mBERT, and xlm-RoBERTa and show that monolingual models perform better than their multilingual counterparts. Their MahaBERT [5] model provides the best results on L3Cube-MahaHate Corpus. In the paper [26], They present results from several machine learning experiments on MOLD[4] dataset, including zero-short and other transfer learning experiments on state-of-the-art cross-lingual transformers from Bengali, English, and Hindi data. Authors [27] release a Marathi dataset and experiment with several machine learning models, including state-of-the-art transformer models, to predict the type and target of offensive tweets in Marathi. Later, attempt using cross-lingual embeddings and transfer learning to spot offensive language. Finally, they investigate semi-supervised data augmentation. They built a larger semi-supervised dataset for Marathi called SeMOLD, which has about 8000 examples.

3. Experimental setup

3.1. Task description

The brief of the task[5]is outlined below.

- **Task 1 ICHCL Binary Classification:** It is ICHCL HINGLISH and GERMAN Codemix Binary Classification. This task aims to identify Hinglish and German hate speech and offensive language. It is a coarse-grained binary classification to classify tweets into two classes: hate and offensive (HOF) and non-hate and offensive (NOT).
  - **(NOT) Non-Hate-Offensive -** This post does not contain hate speech or profane, offensive content.
  - **(HOF) Hate and Offensive -** This post contains hate, offensive, and profane content.
- **Task 2 ICHCL Multiclass CLassification:** Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) - Multiclass Classification.
  This year for the Hinglish language, a multiclass task has been introduced that further divides the HOF tweets into 3 subclasses:
  - **(SHOF) Standalone Hate -** Offensive, profane content is in tweets, comments, or replies.
  - **(CHOF) Contextual Hate -** Comment or reply supporting the hate, offence and profanity expressed in its parent. This includes affirming the hate with positive sentiment and having apparent hate.
  - **(NONE) Non-Hate -** This tweet, comment, or reply does not contain Hate, offensive, or profane content.
- **Task 3A Marathi:** Offensive Language Detection
  - **OFF -** Posts containing any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct.

[4]MOLD is available at: https://github.com/ tharindudr/MOLD

[5]https://hasocfire.github.io/hasoc/2022/call_for_participation.html

  - **NOT -** Posts that do not contain offence or profanity.
- **Task 3B Marathi:** Categorisation of Offensive Language
  - **Targeted Insult (TIN) -** Posts containing an insult/threat to an individual, group, or others.
  - **Untargeted (UNT) -** Posts containing nontargeted profanity and swearing.
- **Task 3C Marathi:** Offense Target Identification
  - **Individual (IND) -** Posts targeting an individual.
  - **Group (GRP) -** The target of these offensive posts is a group of people considered unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristics.
  - **Other (OTH) -** The target of these offensive posts does not belong to any of the previous two categories.

3.2. Dataset

This year, HASOC 2022 provides code-mixed Hinglish-German datasets tagged as "NOT" and "HOF" for binary classification (Task 1) as well as "NONE ", "SHOF" and "CHOF" for multi-classification (Task 2).

Table 1, 2 shows all five task dataset statistics separately. Here, we only include the total count of the test data, not the label count of the test dataset, as it is not provided yet.

Class label	Training	Test
Task 1		
NOT	2,609	-
HOF	2,612	-
TOTAL	5,221	1,077
Task 2		
NONE	2,390	-
SHOF	1,636	-
CHOF	888	-
TOTAL	4,833	996

Table 1
Class distribution analysis for Task 1 and Task 2 dataset, which includes Hinglish-German data

Marathi dataset tagged as "NOT" and "HOF" for binary classification (Task 3A); "NOT", "TIN" and "UNT" for multi-classification (Task 3B); and "NOT", "IND", "GRP" and "OTH" for another multi-classification (Task 3C).

3.3. Preprocessing

- **Data Augmentation:** Here, we utilize the previous year's HASOC-ICHCL2021 data for the binary classification along with the HASOC-ICHCL2022 dataset. We just merged both of the datasets.

Class label	Training	Test
Task 3A		
NOT	2,034	-
HOF	1,069	-
TOTAL	3,103	508
Task 3B		
NOT	2,035	-
TIN	741	-
UNT	327	-
TOTAL	3,103	508
Task 3C		
NOT	2,363	-
IND	503	-
GRP	157	-
OTH	80	-
TOTAL	3,103	508

Table 2
Class distribution analysis for Task 3A, Task 3B and Task 3C dataset which includes only Marathi data

- **Data concatenation:** In preprocessing step, we concatenate tweets, comments, and replies applying the given code[6]. This part is applicable for Task 1 and Task 2.
- **Convert all the words in lowercase:** We convert all the words into lowercase.
- **Converted emojis:** Here, we didn't remove the emoji entirely; rather converted emojis and emoticons to English text [7] as it is a Hinglish code-mix task.
- **Stopwords removal:** We remove English and Hindi stopwords from the dataset.
- **Stemming:** Stemming is used to convert the word to its root word by removing its inflections.
- **Removing unnecessary symbols and url:** Remove @, , *, # , https?:// etc. from the dataset to make the dataset noise free. Applicable for Marathi data also. .
- **Label encoding:** We encode **Class** into a unique number for each task.
  - **Task 1 (HASOC-ICHCL-Hinglish-German2022 binary classification)** - "HOF" to "0", and "NOT" to "1",
  - **Task 2 (HASOC-ICHCL-Hinglish2022 multiclass classification)** - "NONE" to "0", "SHOF" to "1", "CHOF" to "2".
  - **Subtask-3A (HASOC-Marathi2022 binary classification)** "NOT" to "0" and "HOF" to "1".

[6] https://github.com/hasocfire/ICHCLbaseline/tree/master/ICHCL_baseline2k22
[7] https://studymachinelearning.com/text-preprocessing-handle-emoji-emoticon/

– **Subtask-3B (HASOC-Marathi2022 ternary classification)** “NOT” to “0”, “TIN” to “1” and “UNT” to “2”.
– **Subtask-3C (HASOC-Marathi2022 four classification)** “NOT” to “0”, “IND” to “1”, “GRP” to “2” and “OTH” to “3”.

Table 3 shows all the preprocessing steps applied to all five tasks.

Preprocessing steps	Task 1	Task 2	Task 3A	Task 3B	Task 3C
Data Augmentation	Yes	No	No	No	No
Data concatenation	Yes	Yes	No	No	No
Convert all the words in lowercase	Yes	Yes	Yes	Yes	Yes
Converted emojis	Yes	Yes	No	No	No
Stopwords removal	Yes	Yes	No	No	No
Stemming	Yes	Yes	No	No	No
Removing unnecessary symbols and url	Yes	Yes	Yes	Yes	Yes
Label encoding	Yes	Yes	Yes	Yes	Yes

Table 3
Preprocessing steps for all five tasks

3.4. Pre-trained BERT models

BERT models are trained on a large raw text (without human labeling) corpus in a self-supervised way. Figure 2 shows the representation of the general proposed approach for all five tasks.

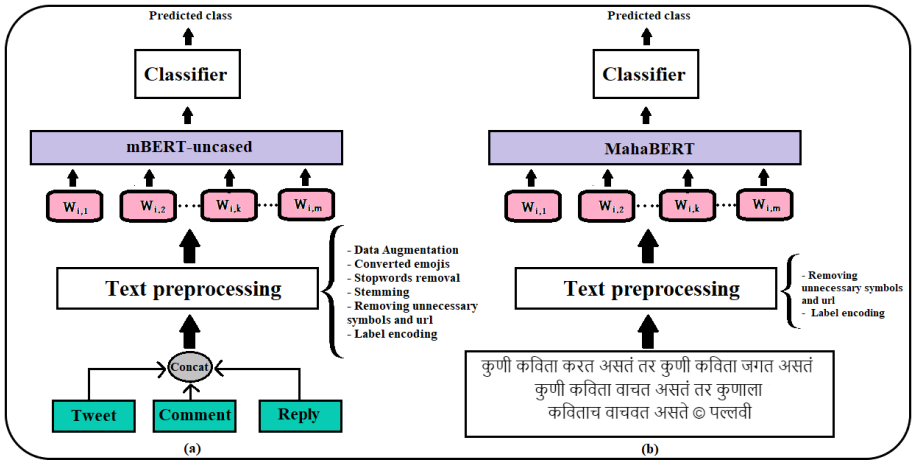

Figure 2: (a) General architecture to perform Task 1 and Task 2 (b) architecture to perform Subtask-3A, Subtask-3B and Subtask-3C

- **mBERT**[8]**:** It is pre-trained with the largest Wikipedia over 104 top languages worldwide, including Hindi, Bengali and Marathi, using a masked language modeling (MLM) objective.

[8] https://huggingface.co/bert-base-multilingual-uncased

For Task-1 and Task-2, we use the same mBERT architecture with a few changes (different preprocessing steps only).

- **MahaBERT[9]:** MahaBERT is a multilingual BERT (bert-base-multilingual-cased) model finetuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets.

For Subtask-3A, Subtask-3B and Subtask-3C, we use the same MahaBERT architecture.

Due to memory and *GPU* issues, we did several experiments but with the same hyperparameter combination (Table 3), and we noticed that smaller batch sizes help better fine-tuning.

Hyperparameter	
Learning-rate	1e-5
Epochs	5
Max seq length	512
Batch size	5

Table 4
Combination of hyperparameters for fine-tuning pre-trained BERT variants

4. Result

Here, table 5 shows the result; Macro F1_Score, precision and recall measures the performance. We put all the tasks' results as shown on the leaderboard. We train the whole dataset and predict classes for the given test set. We also tested other pre-trained BERT models but submitted only one run, giving the best result (we didn't submit other runs as they did not perform well).

Task	**f1_score**	**precision**	**recall**
Task 1 ICHCL Binary Classification	0.6621	0.6732	0.6655
Task 2 ICHCL Multiclass CLassification	0.3952	0.4699	0.4199
Task 3A Marathi	0.9019	0.9021	0.9022
Task 3B Marathi	0.3073	0.3405	0.2868
Task 3C Marathi	0.2063	0.2322	0.1960

Table 5
Performance of all the tasks

5. Conclusion

In this paper, five task performances are presented. In Hinglish-German, our task is to classify a *tweet*, *comment*, and *reply* pair is *HOF* or *NOT* (Task 1). The same pair from the dataset conveys *SHOF* or *CHOF* or *NONE* (Task 2). In Marathi, texts are *HOF* or *NOT* (Subtask-3A). In multiclass

[9] https://huggingface.co/l3cube-pune/marathi-bert

classification, text is *NOT*, *TIN* or *UNT* (Subtask-3B). The last task in Marathi is to classify the text in *NOT* or *IND* or *GRP* or *OTH* (Subtask-3C). We utilized several variants of pre-trained BERT models but submitted only one run. We notice a smaller batch size gives a better result than a larger batch size. Converting emojis and emoticons to text help to increase performance. More experiments on preprocessing are needed to increase the models' performance. Here, data augmentation plays a good role; otherwise, we use a common state-of-the-art baseline transformer-based pre-trained BERT model. We applied the same data augmentation approach for the Marathi dataset, i.e., we merged the previous year's HASOC-Marathi data but couldn't submit it on time; otherwise, it also performed well.

References

[1] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, The British Journal of Criminology 60 (2019) 93–117. URL: https://doi.org/10.1093/bjc/azz049. doi:10.1093/bjc/azz049. arXiv:https://academic.oup.com/bjc/article-pdf/60/1/93/31634412/azz049.pdf.

[2] p. . A. y. M. Satapara, Shrey and Majumder, Prasenjit and Mandl, Thomas and Modha, Sandip and Madhu, Hiren and Ranasinghe, Tharindu and Zampieri, Marcos and North, Kai and Premasiri, Damith, booktitle = FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, ????

[3] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive Language Identification in Marathi, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.

[4] 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.

[5] R. Joshi, L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi BERT language models, and resources, CoRR abs/2202.01159 (2022). URL: https://arxiv.org/abs/2202.01159. arXiv:2202.01159.

[6] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, 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, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[7] M. A. Bashar, R. Nayak, Qutnocturnal@hasoc'19: CNN for hate speech and offensive content identification in hindi language, CoRR abs/2008.12448 (2020). URL: https://arxiv.org/abs/2008.12448. arXiv:2008.12448.

[8] J.-C. Mensonides, P.-A. Jean, A. Tchechmedjiev, S. Harispe, Imt mines ales at hasoc 2019:

automatic hate speech detection, in: FIRE 2019-11th Forum for Information Retrieval Evaluation, volume 2517, 2019, pp. p–279.
[9] S. Mishra, S. Mishra, 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages., in: FIRE (Working Notes), 2019, pp. 208–213.
[10] V. Mujadia, P. Mishra, D. M. Sharma, Iiit-hyderabad at hasoc 2019: Hate speech detection., in: FIRE (Working Notes), 2019, pp. 271–278.
[11] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences 55 (1997) 119–139.
[12] T. Mandl, S. Modha, A. Kumar 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: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.
[13] R. Raj, S. Srivastava, S. Saumya, Nsit & iiitdwd @ hasoc 2020: Deep learning model for hate-speech identification in indo-european languages, in: FIRE, 2020.
[14] S. Kumari, Nohate at hasoc2020: Multilingual hate speech detection, in: Forum for Information Retrieval Evaluation, FIRE, 2020.
[15] 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: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, New York, NY, USA, 2021, p. 1–3. URL: https://doi.org/10.1145/3503162.3503176. doi:10.1145/3503162.3503176.
[16] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, A. Dash, One to rule them all: Towards joint indic language hate speech detection, CoRR abs/2109.13711 (2021). URL: https://arxiv.org/abs/2109.13711. arXiv:2109.13711.
[17] M. Bhardwaj, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Hostility detection dataset in hindi, arXiv preprint arXiv:2011.03588 (2020).
[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, Working Notes of FIRE (2021).
[19] Z. M. Farooqi, S. Ghosh, R. R. Shah, Leveraging transformers for hate speech detection in conversational code-mixed tweets, arXiv preprint arXiv:2112.09986 (2021).
[20] 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, arXiv preprint arXiv:2111.13974 (2021).
[21] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), arXiv preprint arXiv:1903.08983 (2019).
[22] M. Nene, K. North, T. Ranasinghe, M. Zampieri, Transformer models for offensive language identification in marathi, in: FIRE, 2021.
[23] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).

[24] A. Glazkova, M. Kadantsev, M. Glazkov, Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi, arXiv preprint arXiv:2110.12687 (2021).

[25] A. Velankar, H. Patil, A. Gore, S. Salunke, R. Joshi, L3cube-mahahate: A tweet-based marathi hate speech detection dataset and BERT models, CoRR abs/2203.13778 (2022). URL: https://doi.org/10.48550/arXiv.2203.13778. doi:10.48550/arXiv.2203.13778. arXiv:2203.13778.

[26] S. S. Gaikwad, T. Ranasinghe, M. Zampieri, C. Homan, Cross-lingual offensive language identification for low resource languages: The case of Marathi, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 437–443. URL: https://aclanthology.org/2021.ranlp-1.50.

[27] M. Zampieri, T. Ranasinghe, M. Chaudhari, S. Gaikwad, P. Krishna, M. Nene, S. Paygude, Predicting the type and target of offensive social media posts in marathi, Social Network Analysis and Mining 12 (2022) 77. URL: https://doi.org/10.1007/s13278-022-00906-8. doi:10.1007/s13278-022-00906-8.