Hate and Offensive language detection using BERT for English Subtask A

Md Saroar Jahan[1], Djamila Romaissa Beddiar[1], Mourad Oussalah[1], Nabil Arhab[1] and yazid bounab[1]

[1] University of Oulu, Faculty of Information Tech., CMVS, PO Box 4500, Oulu 90014, FINLAND

Abstract
This paper presents the results and main findings of the HASOC-2021 Hate/Offensive Language Identification Subtask A. The work consisted of fine-tuning pre-trained transformer networks such as BERT and an ensemble of different models, including CNN and BERT. We have used the HASOC-2021 English 3.8k annotated twitter dataset. We compare current pre-trained transformer networks with and without Masked-Language-Modelling (MLM) fine-tuning on their performance for offensive language detection. Among different BERT MLM fine-tuned BERT-base, BERT-large, and ALBERT outperformed other models; however, BERT and CNN ensemble classifier that applies majority voting outperformed other models, achieving 85.1% F1 score on both hate/non-hate labels. Our final submission achieved 77.0 F1 in the HASOC-2021 competition.

Keywords
BERT fine-tuning, Offensive language identification, Hate speech, BERT performance comparison.

1. Introduction

The emergence of Web 2.0 platforms that enabled user-generated content and participatory culture has witnessed the proliferation of online hate speech at an unprecedented level, increasing the likelihood of random people of any age group being subject to online harassment and abuse through some internet forum message board or social network platform. Hate speech is a complex phenomenon, intrinsically associated with relationships between groups, and relies on language nuances. Nobata et al. [1] define hate speech as - "Language which attacks or demeans a group based on race, ethnic origin, religion, gender, age, disability, or sexual orientation/gender identity".

In the past few years, the automatic detection of hate speech, cyber-bullying, or aggressive and offensive language became a vividly studied task in natural language processing (NLP). Past research has examined various characteristics of offensive language such as the cyber aggression [2, 3], abusive language [1, 4], hate speech [5, 6], Racism [7] and offensive language [8, 9].

Several workshops (e.g., SemEval-2019[10], SemEval-2020[11], HASOC-2019[12], HASOC-2020[13]) have been organized to find the state-of-the-art practices and new solutions for

efficient offensive text identification.

For example, in SemEval-2019, Task A (offensive language detection) was the most popular sub-task with 104 participating teams. Among the top-performing team Liu et al. [14] used BERT-base-uncased with default-parameters, with a max sentence length of 64 and trained for 2 epochs and achieved 82.9% F1 score. The top nonBERT model by Mahata et al. [15] was ranked fifth. They used an ensemble of CNN and BLSTM+BGRU.

In SemEval-20, 145 teams submitted official runs. The best team Wiedemann et al. [16] achieved an F1 score of 0.9204 using an ensemble of ALBERT models of different sizes. The top-10 teams were close to each other and employed BERT, RoBERTa or XLM-RoBERTa models; sometimes CNNs and LSTMs were also mentioned either for comparison or hybridization purposes.

In HASOC-2020, over 40 research groups participated in HASOC-2020 competition. The top-ranked submission for Hindi-hate speech detection used a CNN with FastText embeddings as input [17]. The best performance for German hate speech detection task was achieved using a fine-tuned versions of BERT, DistilBERT and RoBERTa [18]. Similarly, the top performance in English-language hate speech detection was based on BERT and another deep learning-based model.

This year 2021, HASOC[19] [20] offers three different tasks and a separate dataset for each subtask. Subtask-1A offers tasks in English, Hindi with 2 problems, and Marathi with 1 problem. The subtasks-1B dataset contains English, Hindi, and subtask2 code-mixed Hindi tweets. In our participation, we have participated in subtask-A for English identification of Hate/offensive Twitter posts, and used the HASOC-2021 provided a dataset for training and validation. Regarding the state-of-art practice[21] in the field of hate-speech text detection, our contribution in this paper is threefold:

1. We compare different pre-trained transformer-based neural network model's performance and explain model performance.
2. We study how an additional fine-tuning step with masked language modeling (MLM) of the best individual model conducted on in-domain data affects the model performance.
3. An ensemble of the different models presented, including CNN+BERT.

The paper is structured as follows. In Section 2, we describe our methodology (dataset annotation schema, preprocessing, and classifier architectures including the machine learning models and the associated feature engineering), In Section 3, the details of our result are reported. In Section 4, an error analysis is carried out. Finally, conclusive statements are drawn in Section 5.

2. Methodology

The overall experimentation methodology includes a three-stage process: (i) data collection and preprocessing (ii) experiment machine learning (ML) models architecture, and (iii) error analysis. The experiment environment is the same for all experiments (e.g., data preprocessing, ML architecture, test data, and error analysis). See Figure 1 for a high level description of our methodology whose details are presented in the following subsections.
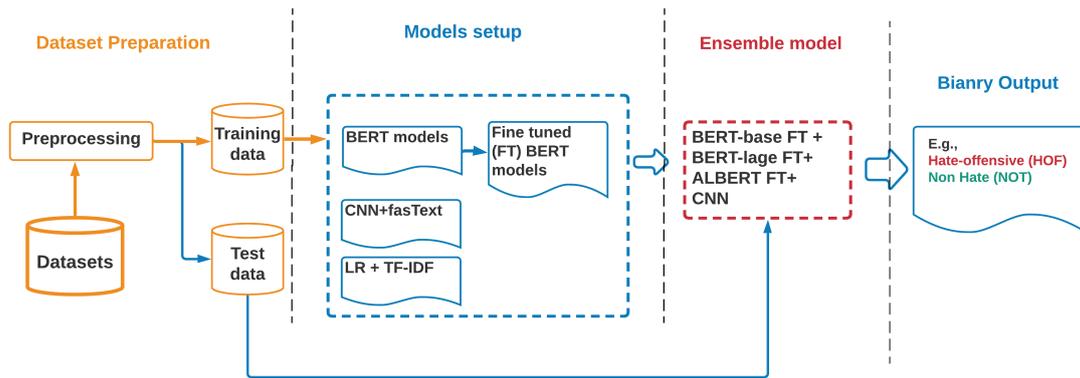
Figure 1: A high level system architecture diagram.

2.1. Dataset

To train our models and compare our results with state-of-the-art models, we used English twitter dataset from HASOC-2021. The HASOC task organizer already annotated datasets for English subtask A. Table 1 shows an example of datasets and annotation. For instance, if the Twitter post contains any hate or offensive word or represents any offensive context, it is considered HOF (hate or/and offensive), otherwise NOT. The total size of the dataset is 3843, among which 2051(65%) contain HOF and the rest 1342 (35%) NOT.

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

Table 1
Example tweets and annotation from English datasets.

Example of Tweet	Task A label	Dataset Size
"@hemantmkpandya @news24tvchannel @Aloksharmaaicc @man-akgupta You are a donkey that's why only one is talking."	HOF	3843
"@For 18-18 hours the termite went and hollowed out a 70 year old strong tree in 7 years !!. #ResignModi"	NOT	
"Fattu hai bjp wala #CruelMamata #BengalViolence #BengalBurning"	HOF	
"Goodbye Sher-e-Bihar… May Allah bless Tala Saheb from high to high in Jannatul Firdous Amen. JusticeForShahabuddin"	NOT	

2.2. Dataset Prepossessing

We have removed special characters, numbers (e.g., @,0-9), newlines, mention tags, and links for data preprocessing. We have not removed hashtags since we found them important for

subsequent reasoning. Table 2 shows an example of preprocessing.

Table 2
Example dataset preprocessing.

Before preprocessing	After preprocessing
@hemantmkpandya @news24tvchannel @Alok-sharmaaicc @manakgupta You are a donkey that's why only one is talking.	You are a donkey that's why only one is talking
Do not look away. #IndiaCovidCrisis https://t.co/oHsnIXlEla	Do not look away. #IndiaCovidCrisis

In order to quantify the influence of the various preprocessing units, we carried out a simple task of HASOC-2021 HOF accuracy rate using Logistic Regression (LR) classifier with TF-IDF features whose results are summarized in Table 3. One can see for instance, that the use of uppercase to lowercase conversion and emoji removal in the preprocessing stage does not affect the overall result. However, Newline + Tab Token, mention tag, and URL + Special Characters removal worked well and improved almost 0.5% in performance accuracy. Since hashtag (#) removal decreases 0.8% performance, we have not removed hashtags from our dataset. This provides a basis for optimal preprocessing pipeline to be used in subsequent tasks.

Table 3
Accuracy scores changes in preprocessing. Result obtain using LR with Tf-IDF.

Preprocessing Type	Accuracy scores
All removed except Hash (#) tag	80.4
URL, Special Ch., Newline, Tab Token	80.1
USERNAME (user) mention tag	80.2
RT	80.1
Emoji	79.9
Stop-word	79.6
Steming	79.4
Lowercase conversion	79.9
Hash (#) tag	78.6
No Preprocessing	79.9

2.3. Models setup

We have used a set of well acknowledged models in hate speech detection tasks as per previous HOF competitions. Especially, three types of classifiers have been utilized: BERT models, CNN and baseline LR model with TF-IDF features as follows:

2.3.1. Convolution Neural Network (CNN) Model

we adopted [22] CNN, architecture, where the input layer is represented as a concatenation of the words forming the post (up to 70 words), except that each word is now represented by its

FastText embedding representation with 300 embedding vectors. A convolution 1D operation with a kernel size 3 was used together with a max-over-time pooling operation over the feature map with a layer dense 50. Dropout on the penultimate layer with a constraint on l2-norms of the weight vector was used for regularization. The details of the implementation are reported on our GitHub page of this project with datasets and codes[1].

2.3.2. Transformer Network Models

BERT – is the Bidirectional Encoder Representations from Transformers: this seminal transformer-based language model applies an attention mechanism that enables learning contextual relations between words in a text sequence [23] (Devlin et al., 2019). Two training strategies that BERT follows:

1. MLM : where 15 % of the tokens in a sequence replaced (masked) for which the model learns to predict the original tokens, and
2. Next sentence prediction (NSP): here the model receives two sentences as input and model learns whether the second sentence is a successor of the first sentence in their original document context.

RoBERTa – is a replication of BERT developed by Facebook [24] with known as Robustly Optimized BERT Pretraining Approach with the following modifications:

1. training the model longer with bigger batches as well as more and cleaner data and discard the NSP objective,
2. training on longer sequences, and
3. dynamically changes the masking patterns, e.g. taking care of masking complete multi-word units. .

XLM-RoBERTa – XLM-R: this is a cross-lingual or multilingual version of RoBERTa which is trained on more than 100 languages at once [25] (Conneau et al., 2019).

ALBERT – represent A Lite BERT, which is a alteration on BERT especially to overcome training time and memory limitations issues [14] (Lan et al., 2019): The main contributions that ALBERT makes over BERT are:

1. decomposing the embedding parameters into smaller matrices that will be projected to the hidden space separately,
2. in contrast to BERT's simpler NSP objective it based on sentence order prediction (SOP), share parameters across layers to improve or stabilize the learned parameters.

2.3.3. Ensemble model

We created an ensemble model using majority voting rule. Especially, we tested ensemble model of i) All BERT models; ii) BERT-large-uncased + BERT-base-uncased + ALBERT-xxlarge-v2; iii) CNN + BERT-large-uncased + BERT-base-uncased + ALBERT-xxlarge-v2. In the case of an even number of models, the ensemble model takes into account the decision weight generated by each classifier to yield the final output (HOF versus NOT).

---

[1]https://github.com/saroarjahan/HASOC-2021-TASKA (accessed September 08, 2021)

3. Experiment Setup

Initially, we employed a random split of the original dataset into 80% for training and 20% for testing and validation, ensuring the same proportion of dataset for all kinds of model learning. Four types of classifiers were implemented: Logistics regression (LR) with word-level TF-IDF, Convolution Neural Network (CNN) with FastText word embedding, and BERT pre-trained model. For BERT model setup, we fine-tuned different transformer models with the HASOC-2021 training data using the corresponding test data for validation. The following models were tested: BERT-base and BERT-large (uncased), RoBERTa-base and RoBERTa-large, XLM-RoBERTa, BERT-multilingual, and four different ALBERT models (large-v1, large-v2, xxlarge-v1, and xxlarge-v2). Each model was fine-tuned for 6 epochs with a learning rate of 5e-6, maximum sequence length of 128, and batch size 4. After each epoch, the model was evaluated on the validation set. The best-performing epoch was saved for the ensembling. We tested ensemble models by majority vote from all models such as BERT-base, BERT-large, and ALBERT.

4. Results

Table 4 shows the results of binary offensive language detection (assuming all tweets as either hate or non-hate) using LR baseline, CNN with word embedding, as well as for the individual fine-tuned transformer models and their corresponding ensembles. CNN, BERT-base-uncased, BERT-large-uncased, and ALBERT-xxlarge-v2 transformer models largely outperform the LR baseline. Comparing BERT and CNN models, BERT-based-uncased and BERT-large-uncased slightly (.1%) outperform CNN model. CNN exhibits a much better performance compared to baseline and some BERT models. achieving 83.3% F1 score.

Our best individual model is BERT-based-uncased with an F1-score of 83.4 %. The experiment also showed that BERT-uncased performed 2% better than BERT-cased. When comparing the different pre-trained transformer models, interesting results emerged as well. For example, none of the Multilingual BERT models has outperformed the BERT-large or BERT-base model. This was not fully a surprise since the multilingual pre-trained model was trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective; however, it contains only a small percentage of English tokens. Therefore, it might have fallen short when the HASOC-2021 dataset is mono English.

We expected dehateBERT would perform better since it pertained only to English and included several hate corpus; however, the results showed only 79.4% F1 scores. One possible explanation could due to our dataset's diverse pattern. Since this dataset only focuses on the Twitter dataset and all the tweets collected from recent tweets and recent events (e.g., Indian politics, Covid). Therefore, dehateBERt cannot generalize much since it was only pre-trained with some specific hate dataset, limiting the dimension of pre-trained data.

Regarding the ensembles of model variants, we can see that the performance is very marginal when all transformers are ensembled together. This is probably due to the fact that some transformers model have not performed well at first hand, so mixing low performed transformed models have reduced the overall performance. However, when we ensembled only the best performing models (BERT-large, BERT-base, and ALBERT), we see our models have increased

Table 4
Accuracy and F1 scores for Hate-Offensive detection Task A, English dataset (best in bold).

Classifier	NOT		HOF		BOTH label	
Feature Name	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
LR+Word Level TF-IDF	69.3	69	92.4	92.2	80.4	80
CNN +Word Emb.	72.1	72	95.4	95.2	83.3	**83.3**
BERT-large-cased	71.2	71	94.3	94.2	82.45	82.15
BERT-large-uncased	72.2	72	95.4	95.3	82.7	**83.4**
BERT-base-cased	69.3	69.1	92.4	92.3	80.3	80
BERT-base-uncased	72.3	72.1	95.6	95.4	83.8	**83.5**
BERT-base-multilingual	65.6	65	88.	88.1	75.7.4	75.3
BERT-base-multilingual-uncased	65.8	65.2	88.7	88.5	76.5	76
RoBERTa-base	65.9	65.7	89	88.9	78.8	78
RoBERTa-large	66.5	66	89.6	89.1	79.5	79
XLM-RoBERTa	65.7	65.3	88.6	88.4	76.6	76.3
ALBERT-large-v1	69.4	69.3	92.5	92.4	80.5	80.1
ALBERT-xxlarge-v2	70.4	69.7	92.6	92.8	81.5	81.3
ALBERT-xxlarge-v2	70	69.3	93	92.7	82.8	**82.4**
Dehatebert-mono-english	69.5	69.2	92.5	92.2	80	79.4
Ensembles of Models (Majority voting applied)						
All BERT model	69.5	69.4	92.9	92.7	81	80.3
BERT-large-uncased + BERT-base-uncased + ALBERT-xxlarge-v2	74.8	74.7	96.1	96	84.9	84.6
CNN + BERT-large-uncased + BERT-base-uncased + ALBERT-xxlarge-v2	76.1	75.8	96.6	96.4	85.5	**85.1**

1% F1 scores. Furthermore, when we ensembled CNN with best performing BERT models, the performance further improved.

We have submitted our best ensemble (CNN + BERT-large-uncased + BERT-base-uncased + ALBERT-xxlarge-v2) results to the HASOC-21 competition, and official result, we have received an 77% F1 score (Table 5).

Table 5
Official results of our HASOC-21 test set submissions for English tasks A.

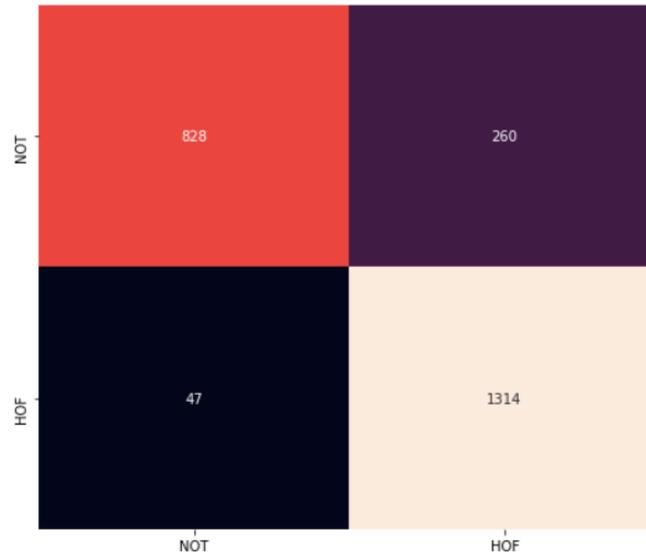| Team name	Task	Macro F1
TeamOulu	English Task A	77

Figure 2: Confusion matrix from 2117 test samples.

5. Error Analysis

Although we have obtained 85.5% accuracy and 85.1% F1 score, the model still exhibits a portion of false detection. To understand this phenomenon better, we performed an error analysis of the model's performance. For this purpose, we randomly prepared subsets of test data, then manually inspected the classifier output. Error test data contained 200 samples, among which 68 were non-hate speech and 132 were hate samples.

From figure 2, we see that most of the error is related to false positive (FP), where our classifier is not performing much while detecting non-hate samples. In contrast, 1414 hate samples were correctly detected, and only 47 resulted as false negative (FN). Since our train dataset majority (65%) contains hate samples, it seems that our model is better trained or biased towards hate-offensive class. Table 6 represents FP and FN metrics. Here, we can see that some data were actually incorrectly annotated (4 false predictions, 3 were wrongly annotated). Since this test data was a split of original data, it indicates that our model performs much better, some false predictions are actually cast into correct prediction, and some of the error comes from the dataset annotation itself.

Table 6
False-positive (FP) and false-negative (FN) examples from first 200 test samples.

Prediction	Tweets	Original Label	Predict label
FP	nah i mean i'm not trying to be a dick here, i'm just trying to make the point that it's not as easy as it 'feels', you know?	0	1
FP	sadly enough innocent poor people of are dying but ur is evolves back from #human to #monky	0	1
FP	you was right dr and you are right ..very sad for our country #murderer_modi #resignmodi	0	1
FN	do i want to die or do i just want to stop feeling empty everyday	1	0

6. Conclusion

After last year's HASOC-2020 shared task on offensive language detection, BERT models emerged as state-of-the-art in HOF, although a fine-tuning of the models is still challenging and open to debate. This motivates our approach in this paper where ensembling state-of-the-art BERT models and CNN, while seeking optimal preprocessing strategy is promoted. Especially, in 2021 competition, we performed different experiments of twitter preprocessing, BERT-fine tune models, and an ensemble of other models. Our tweet preprocessing showed removing the mentioned tag and removing special characters and URLs were useful and increased almost 1% of models' performance. However, eliminating hashtags and stemming reduced the performance of the model. Among different transformers, BERT-base outperformed other models, including BERT-multilingual, RoBERTA, and Dehatebert-mono-english. Among BERT-cased and uncased versions, BERT-uncased showed better performance compared to BERT-cased. Surprisingly, CNN performed much better than most BERT models and performed close to the best performing BERT model. In addition, the ensemble of best-performing models showed further improvement. Our best test ensemble models showed 85.1 F1 scores, and the official submission result showed 77% f1 scores.

7. Acknowledgments

References

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[2] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1–11.

[3] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.

[4] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on arabic social media, in: Proceedings of the first workshop on abusive language online, 2017, pp. 52–56.

[5] Y. J. Foong, M. Oussalah, Cyberbullying system detection and analysis, in: 2017 European Intelligence and Security Informatics Conference (EISIC), IEEE, 2017, pp. 40–46.

[6] C. Abderrouaf, M. Oussalah, On online hate speech detection. effects of negated data construction, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 5595–5602.

[7] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 27, 2013.

[8] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language (2018).

[9] M. S. Jahan, Team oulu at semeval-2020 task 12: Multilingual identification of offensive language, type and target of twitter post using translated datasets, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1628–1637.

[10] 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).

[11] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), arXiv preprint arXiv:2006.07235 (2020).

[12] 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, 2019, pp. 14–17.

[13] 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, 2020, pp. 29–32.

[14] P. Liu, W. Li, L. Zou, Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 87–91.

[15] D. Mahata, H. Zhang, K. Uppal, Y. Kumar, R. Shah, S. Shahid, L. Mehnaz, S. Anand, Midas at semeval-2019 task 6: Identifying offensive posts and targeted offense from twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 683–690.

[16] G. Wiedemann, S. M. Yimam, C. Biemann, Uhh-lt & lt2 at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection, arXiv preprint arXiv:2004.11493 (2020).

[17] R. Raja, S. Srivastavab, S. Saumyac, Nsit & iiitdwd@ hasoc 2020: Deep learning model for

hate-speech identification in indo-european languages (2021).

[18] B. L. R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, Comma@ fire 2020: Exploring multilingual joint training across different classification tasks, in: Working Notes of FIRE 2020-Forum for Information Retrieval Evaluation, Hyderabad, India, 2020.

[19] 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/.

[20] 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: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[21] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, arXiv preprint arXiv:2106.00742 (2021).

[22] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).