Fine-tuning of Pre-trained Transformers for Hate, Offensive, and Profane Content Detection in English and Marathi

Anna Glazkova[1], Michael Kadantsev[2] and Maksim Glazkov[3]

[1]University of Tyumen, 6 Volodarskogo St, Tyumen, 625003, Russian Federation
[2]Thales Canada, Transportation Solutions, 105 Moatfield Dr., Toronto, Canada, M3B 0A4
[3]Neuro.net, 6/16 Alekseevskaya St, Nizhny Novgorod, 603000, Russian Federation

Abstract

This paper describes neural models developed for the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages Shared Task 2021. Our team called *neuro-utmn-thales* participated in two tasks on binary and fine-grained classification of English tweets that contain hate, offensive, and profane content (English Subtasks A & B) and one task on identification of problematic content in Marathi (Marathi Subtask A). For English subtasks, we investigate the impact of additional corpora for hate speech detection to fine-tune transformer models. We also apply a one-vs-rest approach based on Twitter-RoBERTa to discrimination between hate, profane and offensive posts. Our models ranked third in English Subtask A with the F1-score of 81.99% and ranked second in English Subtask B with the F1-score of 65.77%. For the Marathi tasks, we propose a system based on the Language-Agnostic BERT Sentence Embedding (LaBSE). This model achieved the second result in Marathi Subtask A obtaining an F1 of 88.08%.

Keywords

Hate speech, offensive language identification, text classification, transformer neural networks, Twitter-RoBERTa, LaBSE, Marathi

1. Introduction

Social media has a greater impact on our society. Social networks give us almost limitless freedom of speech and contribute to the rapid dissemination of information. However, these positive properties often lead to unhealthy usage of social media. Thus, hate speech spreading affects users' psychological state, promotes violence, and reinforces hateful sentiments [1, 2]. This problem attracts many scholars to apply modern technologies in order to make social media safer. The Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages Shared Task (HASOC) 2021 [3] aims to compare and analyze existing approaches to identifying hate speech not only for English, but also for other languages. It focused on detecting hate, offensive, and profane content in tweets, and offering six subtasks. We participated in three of them:

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
a.v.glazkova@utmn.ru (A. Glazkova); michael.kadantsev@gmail.com (M. Kadantsev); my.eye.off@gmail.com (M. Glazkov)
0000-0001-8409-6457 (A. Glazkova); 0000-0002-2441-2662 (M. Kadantsev); 0000–0002–4290–2059 (M. Glazkov)

- **English Subtask A:** identifying hate, offensive, and profane content from the post in English [4].
- **English Subtask B:** discrimination between hate, profane, and offensive posts in English.
- **Marathi Subtask A:** identifying hate, offensive, and profane content from the post in Marathi [5].

The source code for our models is freely available[1].

The paper is organized as follows. Section 2 contains a brief review of related works. Next, we describe our experiments on the binary and fine-grained classification of English tweets in Section 3. In Section 4, we present our model for hate, offensive, and profane language identification in Marathi. We conclude this paper in Section 5. Finally, Section 6 contains acknowledgments.

2. Related Works

We briefly discuss works done related to harmful content detection in the past few years. Shared tasks related to hate speech and offensive language detection from tweets was organized as a part of some workshops and conferences, such as FIRE [6, 7], SemEval, [8, 9], GermEval [10, 11], IberLEF [12], and OSACT [13]. The participants proposed a broad range of approaches from traditional machine learning techniques (for example, Support Vector Machines [14, 15], Random Forest [16]) to various neural architectures (Convolutional Neural Networks, CNN [17]; Long Short Term Memory, LSTM [18, 19]; Embeddings from Language Models, ELMo [20]; and Bidirectional Encoder Representations from Transformers, BERT [21, 22]). In most cases, BERT-based systems outperformed other approaches.

Most research on hate speech detection continues to be based on English corpora. Despite this, the harmful content is distributed in different languages. Therefore, there have been previous attempts at creating corpora and developing models for hate speech detection in common non-English languages, such as Arabic [13, 23], German [6, 7, 10, 11], Italian [24, 25], Spanish [9, 12], Hindi [6, 7], Tamil and Malayalam [7]. Several studies have focused on collecting hate speech corpora for Chinese [26], Portuguese [27], Polish [28], Turkish [29] and Russian [30] languages.

3. English Subtasks A & B: Identification and Fine-grained Classification of Hate, Offensive, and Profane Tweets

The objective of English Subtasks A & B is to identify whether a tweet in English contains harmful content (Subtask A) and perform a fine-grained classification of posts into three categories, including: hate, offensive, or profane (Subtask B).

[1] https://github.com/ixomaxip/hasoc

Table 1
Data description.

Label	Description	Number of examples (training set)
Subtask A		
NOT	Non Hate-Offensive: the post does not contain hate speech, profane, offensive content	1102
HOF	Hate and Offensive: the post contains hate, offensive, or profane content.	1972
Subtask B		
NONE	The post does not contain hate speech, profane, offensive content	1102
HATE	Hate speech: the post contains hate speech content.	542
OFFN	Offensive: the post contains offensive content.	482
PRFN	Profane: the post contains profane words.	948

3.1. Data

The dataset provided to the participants of the shared task contains 4355 manually annotated social media posts divided into training (3074) and test (1281) sets. Table 1 presents the data description.

Further, we tested several data sampling techniques using different hate speech corpora as additional training data. Firstly, we evaluated the joint use of multilingual data provided by the organizers of HASOC 2021, including the English, the Hindi, and the Marathi training sets. Secondly, as the training sets were highly imbalanced, we applied the positive class random oversampling technique so that each training batch contained approximately the same number of samples. Besides, we experimented with the seq2seq-based data augmentation technique [31]. For this purpose, we fine-tuned the BART-base model for the denoising reconstruction task where 40% of tokens are masked and the goal of the decoder is to reconstruct the original sequence. Since the BART model [32] already contains the <mask> token, we use it to replace mask tokens. We generated one synthetic example for every tweet in the training set. Thus, the augmented data size is the same size as the size of the original training set. Finally, we evaluated the impact of additional training data, including: (a) the English dataset, used at HASOC 2020 [7]; (b) HatebaseTwitter, based on the hate speech lexicon from Hatebase[2] [8]; (c) HatEval, a dataset presented at Semeval-2019 Task 5 [9] ; (d) Offensive Language Identification Dataset (OLID), used in the SemEval-2019 Task 6 (OffensEval) [33]. All corpora except the HatebaseTwitter dataset contain non-intersective classes. Besides, all listed datasets are collected from Twitter. A representative sampling of additional data is shown in Table 2.

We preprocessed the datasets for Subtasks A & B in a similar manner. Inspired by [34], we used the following text preprocessing technique[3]: (a) removed all URLs; (b) replaced all user mentions with the $MENTION$ placeholder.

[2]https://hatebase.org/
[3]https://pypi.org/project/tweet-preprocessor

Table 2
Hate-related dataset characteristics.

Dataset	Size	Labels
HASOC 2020	4522	HOF - 50.4% NOT - 49.6%
HatebaseTwitter	24783	hate speech - 20.15% offensive language - 85.98% neither - 23.77%
HatEval	13000	1 (hate speech) - 42.08% 0 (not hate speech) - 57.92%
OLID	14100	OFF - 32.91% NOT - 67.09%

3.2. Models

We conduct our experiments with neural models based on BERT [35] as they have achieved state-of-the-art results in harmful content detection. For example, BERT-based models proved efficient at previous HASOC shared tasks [7, 6] and SemEval [33, 36].

We used the following models:

- BERT$_{base}$ [35], a pre-trained model on BookCorpus [37] and English Wikipedia using a masked language modeling objective.
- BERTweet$_{base}$ [38], a pre-trained language model for English tweets. The corpus used to pre-train BERTweet consists of 850M English Tweets including 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic.
- Twitter-RoBERTa$_{base}$ for Hate Speech Detection [34], a RoBERTa$_{base}$ [39] model trained on 58M tweets and fine-tuned for hate speech detection with the TweetEval benchmark.
- LaBSE [40], a language-agnostic BERT sentence embedding model supporting 109 languages.

3.3. Experiments

For both Subtask A and Subtask B, we adopted pre-trained models from HuggingFace [41] and fine-tuned them using PyTorch [42]. We fine-tuned each pre-trained language model for 3 epochs with the learning rate of 2e-5 using the AdamW optimizer [43]. We set batch size to 32 and maximum sequence size to 64. To validate our models during the development phase, we divided labelled data using the train and validation split in the ratio 80:20.

Table 3 shows the performance of our models on the validation subset for Subtask A in terms of macro-averaging F1-score (F1), precision (P), and recall (R). As can be seen from the table, BERT, BERTweet, and LaBSE show very close results during validation. Despite this, LaBSE jointly fine-tuned on three mixed multilingual datasets shows the highest precision score. The use of Twitter-RoBERTa increases the F1-score by 1.5-2.5% compared to other classification models. Based on this, we chose Twitter-RoBERTa for further experiments. We found out that neither the random oversampling technique nor the use of the augmented and additional data shows a performance improvement, except the joint use of the original dataset and the

Table 3
Model validation results for English Subtask A, %.

Model	F1	P	R
BERT	79.24	79.74	78.82
BERTweet	78.65	79.36	78.08
Twitter-RoBERTa	81.1	80.01	82.65
LaBSE (English dataset)	78.83	79.5	78.29
LaBSE (English + Hindi)	79.32	79.95	78.8
LaBSE (English, Hindi, and Marathi)	79.27	**81.74**	77.79
Adding extra data to Twitter-RoBERTa			
+ random oversampling	79.97	79.9	80.04
+ BART data augmentation	79.24	78.44	80.31
+ HASOC 2020	78.79	77.66	80.47
+ HatabaseTwitter	**81.19**	79.99	**82.93**
+ HatEval	74.31	75.53	73.64
+ OLID	79.29	78.17	80.93

HatebaseTwitter dataset that gives an F1-score growth of 0.09% and a precision growth of 0.28% compared to basic Twitter-RoBERTa.

For our official submission for Subtask A, we designed a soft-voting ensemble of five Twitter-RoBERTa jointly fine-tuned on the original training set and the HatebaseTwitter dataset (see Table 4). For Subtask B, we used the following one-vs-rest approach to discrimination between hate, profane, and offensive posts.

- First, we applied our Subtask A binary models to identify non hate-offensive examples.
- Second, we fine-tuned three Twitter-RoBERTa binary models to delimit examples of hate-vs-profane, hate-vs-offensive, and offensive-vs-profane classes. The training dataset was extended with the HatebaseTwitter dataset.
- Finally, we compared the results of binary models. If the result was defined uniquely, we used it as a predicted label. Otherwise, we chose the label in proportion to the number of examples in the training set.
  This can be illustrated briefly by the following examples.
  - Let the models show the following results:
    * hate-vs-profane⟶hate;
    * hate-vs-offensive⟶hate;
    * offensive-vs-profane⟶offensive.

    Thus, classes have the following votes: hate – 2, offensive - 1, profane – 0. Then we predict the HATE label.
  - If the results are:
    * hate-vs-profane⟶profane;
    * hate-vs-offensive⟶hate;
    * offensive-vs-profane⟶offensive,

    we have the class votes, such as hate – 1, offensive - 1, profane – 1. Then we choose the PRFN label as the most common label in the training set.

Table 4
Performance of our final models for English Subtasks A & B, official results, %.

Subtask	F1 (our model)	F1 (winning solution)	P (our model)	P (winning solution)	Avg F1	Number of submitted teams	Rank
A	81.99	83.05	84.68	84.14	75.7	56	3
B	65.77	66.57	66.32	66.88	57.07	37	2

4. Marathi Subtask A: Identifying Hate, Offensive, and Profane Content from the Post

4.1. Data

For the Marathi task, we used the original training and test sets provided by the organizers of the HASOC 2021. The whole dataset contains 2499 tweets, including: 1874 training and 625 test examples. The training set consists of 1205 texts of the NOT class and 669 texts of the HOF class. We used raw data as an input for our models. Following [44, 45], we experimented with the combination of the English, the Hindi, and the Marathi training sets provided by the organizers.

4.2. Models

We evaluated the following models:

- XLM-RoBERTa$_{base}$ [46], a transformer-based multilingual masked language model supporting 100 languages.
- LaBSE [40], a language-agnostic BERT sentence embedding model pre-trained on texts in 109 languages.

4.3. Experiments

We experimented with the above-mentioned language models fine-tuned on monolingual and multilingual data. For model evaluation during the development phase, we used the random train and validation split in the ratio 80:20 with a fixed seed. We set the same model parameters as for English tasks.

Table 5 illustrates the results. It can be seen that LaBSE outperforms XLM-RoBERTa in all cases. Moreover, the F1-score of LaBSE fine-tuned only on the Marathi dataset are higher than the results of LaBSE fine-tuned on multilingual data. XLM-RoBERTa, on the other hand, mostly benefits from multilingual fine-tuning.

For our final submission, we used a soft-voting ensemble of five LaBSE fine-tuned on the official Marathi dataset provided by the organizers of the competition. The results of this model on the test set are shown in Table 6.

Table 5
Model validation results for Marathi Subtask A, %.

Model	F1	P	R
XLM-RoBERTa (Marathi dataset)	83.87	85.39	83.39
XLM-RoBERTa (Marathi + Hindi)	83.23	83.82	82.76
XLM-RoBERTa (Marathi + English)	84.83	85.03	84.64
XLM-RoBERTa (Marathi + Hindi + English)	84.35	84.82	83.95
LaBSE (Marathi)	**87.76**	87.82	**87.68**
LaBSE (Marathi + Hindi)	87.62	**88.21**	87.13
LaBSE (Marathi + English)	87.62	**88.21**	87.13
LaBSE (Marathi + Hindi + English)	86.34	86.63	86.08

Table 6
Performance of our final model for the Marathi Subtask A, official results, %.

F1 (our model)	F1 (winning solution)	P (our model)	P (winning solution)	Avg F1	Number of submitted teams	Rank
88.08	91.44	87.58	91.82	82.55	25	2

5. Conclusion

In this paper, we have presented the details about our participation in the HASOC Shared Task 2021. We have explored an application of domain-specific monolingual and multilingual BERT-based models to the tasks of binary and fine-grained classification of Twitter posts. We also proposed a one-vs-rest approach to discrimination between hate, offensive, and profane tweets. Further research can focus on analyzing the effectiveness of various text preprocessing techniques for harmful content detection and exploring how different transfer learning approaches can affect classification performance.

6. Acknowledgments

The work on multi-label text classification was carried out by Anna Glazkova and supported by the grant of the President of the Russian Federation no. MK-637.2020.9.

References

[1] L. E. Beausoleil, Free, hateful, and posted: rethinking first amendment protection of hate speech in a social media world, BCL Rev. 60 (2019) 2101.
[2] M. Bilewicz, W. Soral, Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization, Political Psychology 41 (2020) 3–33.
[3] 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.
[4] 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/.
[5] S. Gaikwad, T. Ranasinghe, M. Zampieri, C. M. Homan, Cross-lingual offensive language identification for low resource languages: The case of Marathi, in: Proceedings of RANLP, 2021.
[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, 2019, pp. 14–17.
[7] 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.
[8] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 11, 2017.
[9] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63.
[10] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language, in: 14th Conference on Natural Language Processing KONVENS 2018, 2018.
[11] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, et al., Overview of GermEval task 2, 2019 shared task on the identification of offensive language (2019).
[12] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of detoxis at IberLEF 2021: Detection of toxicity in comments in Spanish, Procesamiento del Lenguaje Natural 67 (2021) 209–221.
[13] H. Mubarak, K. Darwish, W. Magdy, T. Elsayed, H. Al-Khalifa, Overview of OSACT4 Arabic offensive language detection shared task, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 48–52.
[14] F. Schmid, J. Thielemann, A. Mantwill, J. Xi, D. Labudde, M. Spranger, Fosil-offensive language classification of German tweets combining SVMs and deep learning techniques, in: KONVENS, 2019.
[15] S. Hassan, Y. Samih, H. Mubarak, A. Abdelali, A. Rashed, S. A. Chowdhury, Alt submission for osact shared task on offensive language detection, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 61–65.
[16] B. Ray, A. Garain, JU at HASOC 2020: Deep learning with RoBERTa and random forest for

hate speech and offensive content identification in Indo-European languages, in: FIRE (Working Notes), 2020, pp. 168–174.
[17] A. Ribeiro, N. Silva, Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 420–425.
[18] A. K. Mishraa, S. Saumyab, A. Kumara, IIIT_DWD@ HASOC 2020: Identifying offensive content in Indo-European languages (2020).
[19] A. Montejo-Ráez, S. M. Jiménez-Zafra, M. A. García-Cumbreras, M. C. Díaz-Galiano, SINAI-DL at SemEval-2019 task 5: Recurrent networks and data augmentation by paraphrasing, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 480–483.
[20] M. Bojkovsky, M. Pikuliak, STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on twitter with MUSE and ELMo embeddings, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 464–468.
[21] J. Risch, A. Stoll, M. Ziegele, R. Krestel, hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model., in: KONVENS, 2019.
[22] 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.
[23] N. Albadi, M. Kurdi, S. Mishra, Are they our brothers? analysis and detection of religious hate speech in the Arabic twittersphere, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 69–76.
[24] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the EVALITA 2018 hate speech detection task, in: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 2263, CEUR, 2018, pp. 1–9.
[25] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[26] X. Tang, X. Shen, Y. Wang, Y. Yang, Categorizing offensive language in social networks: A Chinese corpus, systems and an explanation tool, in: China National Conference on Chinese Computational Linguistics, Springer, 2020, pp. 300–315.
[27] R. P. de Pelle, V. P. Moreira, Offensive comments in the Brazilian web: a dataset and baseline results, in: Anais do VI Brazilian Workshop on Social Network Analysis and Mining, SBC, 2017.
[28] M. Ptaszynski, A. Pieciukiewicz, P. Dybała, Results of the PolEval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter (2019).
[29] Ç. Çöltekin, A corpus of Turkish offensive language on social media, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6174–6184.
[30] L. Komalova, A. Glazkova, D. Morozov, R. Epifanov, L. Motovskikh, E. Mayorova, Automated classification of potentially insulting speech acts on social network sites, in: International Conference on Digital Transformation and Global Society, Springer, 2021.
[31] V. Kumar, A. Choudhary, E. Cho, Data augmentation using pre-trained transformer models, in: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems,

2020, pp. 18–26.
[32] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.
[33] 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), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.
[34] F. Barbieri, J. Camacho-Collados, L. E. Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 1644–1650.
[35] 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).
[36] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, SemEval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 59–69.
[37] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.
[38] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[39] 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).
[40] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, arXiv preprint arXiv:2007.01852 (2020).
[41] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019) 8026–8037.
[43] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2018.
[44] S. Mishra, S. Prasad, S. Mishra, Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 120–125.
[45] P. Singh, P. Bhattacharyya, CFILT IIT Bombay at HASOC 2020: Joint multitask learning of multilingual hate speech and offensive content detection system., in: FIRE (Working Notes), 2020, pp. 325–330.
[46] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave,

M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.