Exploring Transformer Based Models to Identify Hate Speech and Offensive Content in English and Indo-Aryan Languages

Somnath Banerjee[1], Maulindu Sarkar[2], Nancy Agrawal[2], Punyajoy Saha[1] and Mithun Das[1]

[1]*Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India*
[2]*Department of Electrical Engineering , Indian Institute of Technology, Kharagpur, West Bengal, India*

Abstract

Hate speech is considered to be one of the major issues currently plaguing online social media. Repeated and repetitive exposure to hate speech has been shown to create physiological effects on the target users. Thus, hate speech, in all its forms, should be addressed on these platforms in order to maintain good health. In this paper, we explored several Transformer based machine learning models for the detection of hate speech and offensive content in English and Indo-Aryan languages at FIRE 2021. We explore several models such as mBERT, XLMR-large, XLMR-base by team name "Super Mario". Our models came $2nd$ position in Code-Mixed Data set (Macro F1: 0.7107), $2nd$ position in Hindi two-class classification (Macro F1: 0.7797), $4th$ in English four-class category (Macro F1: 0.8006) and $12th$ in English two-class category (Macro F1: 0.6447). We have made our code public [1].

Keywords

Hate speech, offensive speech, classification, low resource languages, Hindi, Marathi

1. Introduction

Online Social media platforms such as Twitter, Facebook have connected billions of people and allowed them to publish their ideas and opinions instantly. The problem arises when the bad actors(users) share contents to spread propaganda, fake news, and hate speech etc[1] by using these platforms. Companies like Facebook have been accused of instigating anti-Muslim mob violence in Sri Lanka that left three people dead [1] and a UN report blamed them for playing a leading role in the possible genocide of the Rohingya community in Myanmar by spreading hate speech [2]. In order to mitigate the spread of hateful/offensive content, these platforms have come up with some guidelines[3] and expect that users should follow the guidelines before sharing any

---

[1]https://github.com/hate-alert/OffensiveLangDetectIndoAryan

[1] http://www.aaiusa.org/unprecedented_increase_expected_in_upcoming_fbi_hate_crime_report

[2] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate

[3]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

content. Sometimes, violation of such procedures could lead to the post being deleted or user account suspension.

To reduce the harmful content (such as offensive/hate speech) from these platforms, they employ moderators [2] to keep the conversations healthy and people-friendly by manually checking the posts. With the ever-increasing volume of data on the platform, manual moderation does not seem a feasible solution in the long run. Hence, platforms are looking toward automatic moderation systems for maintaining civility in their platforms. It has already been observed that Facebook has actively removed a large portion of malicious content from their platforms even before the users report them [3]. However, the limitation is these platforms can detect such abusive content in specific major languages, such as English, Spanish, etc. [4, 5]. Hence, an effort is required to detect and mitigate offensive/hate speech-language in low resource language. It has been found that Facebook has the highest number of users, and Twitter has the third-highest number of users in India. So it is necessary for these platforms to have moderation systems for Indian languages as well.

There is a lot of state-of-the-art hate speech detection research content present in the market, mostly in English languages [6]. To extend the research in other languages, we also study these methods, which detect offensive/ hate content in Hindi, Marathi, Code-Mixed languages using the data in this shared task.

Despite being the third most spoken language, Hindi is always being considered a low resource language because of its mostly typological representation. Marathi is also kind of very low resource language because there is rarely some work present to identify Hate/ Offensive content. Finally, Code-mixed data is also following a current trend because of complexity in writing local languages in universal key inputs.

Earlier, in HASOC 2019[4] three datasets have been launched to identify Hate and offensive content in English, German and Hindi languages, and in HASOC 2020[5] another dataset has been launched, aiming to identify offensive posts in the code-mixed dataset. Extending the previous work, this time HASOC[7, 8] has introduced two Sub-task, where Sub-task 1 is further divided into two parts. Sub-task 1A focus on Hate speech and Offensive language identification offered for English, Hindi and Marathi. Sub-task 1A is a coarse-grained binary classification in which the posts have to be classified into two classes, namely: Hate and Offensive (HOF) and Non-Hate and offensive (NOT). Sub-task 1B is a fine-grained classification offered for English and Hindi. Hate-speech and offensive posts from sub-task A are further classified into three categories, Hate speech, offensive and profane. On the other hand, Sub-task 2 [9] focuses on identifying conversational hate-speech in code-mixed languages. In Sub-task 1B, we participated only in the English language. The definitions of different class labels are given below:

- **HATE - Hate Speech [10]**: A post is targeting a specific group of people based on their ethnicity, religious beliefs, geographical belonging, race, etc., with malicious intentions of spreading hate or encouraging violence.

---

[4]https://hasocfire.github.io/hasoc/2019/index.html
[5]https://hasocfire.github.io/hasoc/2020/index.html

- **OFFN - Offensive [11]** [6]: Offensive describes rude or hurtful behaviour or a military or sports incursion into an opponent's territory. In any context, "on the offensive" means on the attack.

- **PRFN - Profane [12]** [7]: A post that expresses deeply offensive behaviour shows a lack of respect, especially for someone's religious beliefs.

In this paper, we have investigated several Transformer based models for our classification task, which has already been seen to be outperforming the existing baselines and standing as a state-of-the-art model. We perform pre-processing, data sampling, hyper-parameter tuning etc., to build the model. Our models are standing in the $2nd$ position in Code-Mixed Data set, $2nd$ position in Hindi two-class classification, $4th$ in English four-class classification and $12th$ in English two-class classification. The rest of the paper is organised as follows: Related literature for Hate speech and offensive language detection Section 2. We have discussed the Dataset Description in Section 3; In Section 4, we have presented the System Description. Finally, we have evaluated the experimental setup in Section 5 and the Conclusion in Section 6.

2. Related Works

The problem of hate/offensive speech has been studied for a long time in the research community. People were continuously trying to improve the models in order to identify hateful/offensive content more precisely. One of the earliest works that tried to detect hate speech by using lexicon-based features [13]. Although they have provided an efficient framework for future research, their dataset was short for any conclusive evidence. In 2017, Davidson et al. [14] contributed a dataset in which thousands of tweets were labelled hate, offensive, and neither. With the classification task of detecting hate/offensive speech present in Tweets in mind. Using this dataset, they then explored how linguistic features such as character and word n-grams affected the performance of a classifier aimed to distinguish the three types of Tweets. Additional features in their classification involved binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet. The authors found that one of the issues with their best performing models was that they could not distinguish between hate and offensive posts. With the advent of neural networks becoming more accessible and usable for people, many of them tried solutions using these models.

In 2018, Pitsilis et al. [15], tried deep learning models such as recurrent neural networks (RNNs) to identify the offensive language in English and found that it was quite effective in this task. RNN's remember the output of each step the model conducts. This approach can capture linguistic context within a text, which is critical to detection. In contrast, RNN's have been projected to work well with language models, other neural network models, such as CNN. LSTM has had notable success in detecting hate/offensive speech [16, 17].

---

[6]https://www.vocabulary.com/dictionary/offensive
[7]https://www.vocabulary.com/dictionary/profane

Although the research on hate/offensive speech detection has been growing rapidly, one of the current issues is that most of the datasets are available in the English language only. Thus, hate/offensive speech in other languages are not detected properly and this could be harmful. This is also a problem for companies like Facebook, which can only detect hate speech in certain languages (English, Spanish, and Mandarin) [4]. Recently, the research community has begun to focus on hate/offensive language detection in other low resourced languages like Danish [18], Greek [19] and Turkish[20]. In the Indian context, the HASOC 2019 [8] shared Task by Mandal et al. [21] was a significant effort in that direction, where authors created a dataset of hateful and offensive posts in Hindi and English. The best model in this competition used an ensemble of multilingual Transformers, fine-tuned on the given dataset [22]. In the Dravidian part of HASOC 2020 [9], Renjit and Idicula [23] used an ensemble of deep learning and simple neural networks to identify offensive posts in Manglish (Malayalam in the roman font).

Recently, Transformer based [24] language models such as BERT, m-BERT, XLM-RoBERTa [25] are becoming quite popular in several downstream tasks, such as classification, spam detection etc. Previously, it has been already seen these Transformer based models have been outperformed several deep learning models [10] such as CNN-GRU, LSTM etc. Having observed the superior performance of these Transformer based models, we focus on building these models for our classification problem.

3. Dataset Description

The shared tasks present in this competition are divided into two parts. The datasets have been sampled from Twitter. Subtask-1 offers in English, Hindi with two problems, and Marathi with one problem. The subtasks-2 dataset contains English, Hindi and code-mixed Hindi tweets. Details about the problem statements have been discussed below:

3.1. Subtask 1A: Identifying Hate, offensive and profane content from the post

The primary focus of Subtask 1A on Hate speech and Offensive language identification, mainly for English, Hindi and Marathi [26], is coarse-grained binary classification. In Table 1 we have presented the dataset statistics on English and Hindi for binary classification.

3.2. Subtask 1B: Discrimination between Hate, profane and offensive posts

This subtask is a fine-grained classification of English and Hindi. Mostly Hate-speech and offensive posts from Subtask 1A are further classified as (HATE) Hate speech, (OFFN) Offensive, (PRFN) Profane, (NONE) Non-Hate. In Table 2 we have presented the dataset statistics on English for Four-class classification (We participated for English language only.).

Figure 1: Example of conversational Hate speech (Image has been taken from HASOC 2021 website)

Category	English		Hindi		Marathi		Code-Mixed	
	Train	Test	Train	Test	Train	Test	Train	Test
(NOT) Non Hate-Offensive	1342	483	3161	1027	1205	418	2899	695
(HOF) Hate and Offensive	2501	798	1433	505	669	207	2841	653
Total	3843	1281	4594	1532	1874	625	5740	1348

Table 1
Two-Class Dataset statistics for languages English, Hindi, Marathi and Code-Mixed

3.3. Subtask 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL)

A conversational thread can also contain hate and offensive content, which not always can be identified from a single comment or reply to a comment. In this type of situation, context is important to identify the hate or offensive content.

According to Figure 1 the parent tweet is expressing hate and profanity towards Muslim countries regarding the controversy happening in Israel at the time. The two comments on the tweet have written "Amine", which means "truthfully" in Persian, which supports the hate but with the context of the parent. This sub-task focused on the binary classification of such conversational tweets with tree-structured data into (NOT) Non-Hate-Offensive and (HOF) Hate and Offensive. In Table 1 we have presented the dataset statistics of the code-mixed data for binary classification as well.

3.4. Pre-Processing

While manually going through the data, we found the dataset contains lots of special characters, emoji's, blank spaces, links etc. Mostly custom functions have been used in pre-processing the

Category	English	
	Train	Test
(HATE) Hate speech	683	224
(OFFN) Offensive	622	195
(PRFN) Profane	1196	379
(NONE) Non-Hate	1342	483
Total	3843	1281

Table 2
Four-Class Dataset statistics for languages English
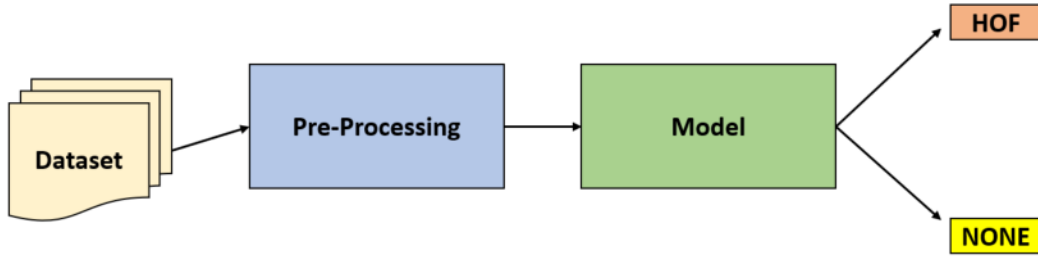
Figure 2: Design of the proposed methods

files, but some libraries were helpful, like "emoji", "nltk" as a baseline. Performed pre-processing steps are:

- We have replaced all the tagged user names to @user.
- We have removed all non-alphanumeric characters except full stop and punctuation (| , ?) in Hindi and Marathi. We have kept all the stop words because by that way machine will be able to identify the sequence of characters properly.
- We have removed emojis, flags and emotions.
- We have removed all the URLs.
- We have kept the hashtags because the hashtags contains some contextual information

4. System Description

We have presented our proposed models for Offensive language detection and Hate speech detection in English, Marathi, Hindi and Code-mixed posts. The overall pipeline of the methodology has been represented in Figure 2. The baseline we have used is Transformer based pre-trained architecture of BERT [25]. More intuitively, we have used a couple of various versions of

BERT, more specifically mBERT [25] and XLM-Roberta [27]. The beauty of XLM-Roberta is it is trained in an unsupervised manner on the multilingual corpus. XLM-Roberta has achieved state-of-the-art results in most language modelling tasks.

4.1. Binary Classification

Most of our task was binary classification problem based on respective embedding. We fine-tuned BERT Transformer and classifier layer on top and used binary target labels for individual classes. We have used this procedure with dehatebert-mono-english [28] and XLM-Roberta [27] for English dataset. We have used multilingual BERT (mBERT) and XLM-Roberta for Marathi, Hindi and Code-Mixed classification. Binary cross-entropy loss can be computed for previously mentioned classification task can be mathematically formulated as:

$$Loss = -\sum_{i=1}^{C'=2} t_i log(s_i) = -t_1 log(s_1) - (1 - t_1) log(1 - s_1)$$

Where it's assumed that there are two classes: C1 and C2. t1 [0,1] and s1 are the ground truth and the score for C1, and t2=1-t1 and s2=1-s1 are the ground truth and the score for C2.

4.2. Multi-class Classification

In this procedure, we have considered the problem as a Multi-Class classification task. We have fine-tuned the BERT model to get the contextualized embedding by the attention mechanism. We have tried Weighted XLM-Roberta large and weighted dehatebert-mono-english for the Four-Class classification task.

4.3. Weighted Binary Classification

The main challenge in any classification problem is the imbalance in data. This imbalance in data may create a bias towards the most present labels, which leads to a decrease in classification performance. According to table 1 it is clearly evident that except code-Mixed data set, there are class imbalances present in the English, Hindi and Marathi dataset. In the English dataset (HOF), Hate and Offensive labels are 46% more than (NOT) Non-Hate-Offensive class. Similarly, in Hindi (NOT), Non-Hate-Offensive class labels are 54% more than (HOF) Hate and Offensive class. In Marathi, also (NOT) Non-Hate-Offensive class labels are 44% more than (HOF) Hate and Offensive class.

There is a lot of research has been done in this domain to make the data balance. Oversampling and Undersampling are very much popular data balancing methods, but they have coherent disadvantages also. We tried to implement data balance by using the class weight procedure. Table 3 describes the class weight distribution we have used in order to manage the data imbalance.

4.4. Weighted Multi-Class Classification

In Multi-Class classification also clearly, there is data imbalance present, and we normalized it. It is evident from Table 3 that (HATE) Hate Speech and (OFFN) Offensive counts are quite

Task Name	(NOT) Non Hate-Offensive Class Weight	(HOF) Hate and Offensive Class Weight
English	1.4318	0.7682
Hindi	0.7266	1.6029
Marathi	0.7775	1.4005

Table 3
Normalized Class Weight for Two Class Classification

Task Name	(HATE) Hate speech Class Weight	(OFFN) Offensive Class Weight	(PRFN) Profane Class Weight	(NONE) Non-Hate Class Weight
English	1.4066	1.5446	0.8033	0.7159

Table 4
Normalized Multi-Class Classweight

similar, but they are almost 50% less than (PRFN) Profane and (NONE) Non-Hate individually. We computed class weight for Multi-Class classification, which is present in Table 4.

4.5. Tuning Parameters

For all the models presented here, we have pre-train on the target dataset for 20 epochs in order to capture the semantics. Along with that, we fine-tuned weighted and unweighted using cross-entropy loss functions[29]. We have used HuggingFace[30] and PyTorch [31]. Initial phases, we have used Adam optimizer[32] with an initial learning rate as 2e-5. We have not used early stopping while training.

5. Results

Our observation was among most of the individual Transformer based BERT models, and the best performance was coming using XLM-Roberta-large (XLMR-large) in English Two-Class and Four-Class dataset and Marathi dataset. In contrast, we are getting the best performance in Code-Mixed dataset by using Custom XLM-Roberta-large. In the case of the Hindi dataset, mBERT is giving the best performance. The beauty of XLM-Roberta is that it has been pre-trained on the parallel corpus. We have noted that the performance of XLM-Roberta-large is very much consistent with most of the regional languages.

While achieving the performance scores. We have used multiple random seeds and have observed that performance was heavily getting impacted for different seeds. It has been observed that while using mBERT, the performance varied 6-7% across our experimented languages. In the case of XLM-Roberta models, the performance was mostly the same and, it varied a maximum 1-2%. Table 5 shows the performance of XLMR-base, XLMR-large and mBERT-base for Two

Classifiers	English	Hindi	Code-Mixed	Marathi
	Macro F1	Macro F1	Macro F1	Macro F1
XLMR-base	0.7834	0.6862	0.6456	0.8133
XLMR-large	**0.8006**	0.7112	**0.7107**	**0.8756**
mBERT-base	0.7328	**0.7797**	0.6277	0.8611
Indic-BERT	0.7002	0.6323	0.5912	0.8176
dehate-BERT	0.7811	0.6533	0.6377	0.7550
Submission Name	"Bestn"	"T2"	"Context 1"	-

Table 5
Two-Class Classification Result

Classifiers	English
	Macro F1
XLMR-base	0.5824
XLMR-large	**0.6447**
mBERT-base	0.5443
Indic-BERT	0.5119
dehate-BERT	0.4845
Submission Name	"Final"

Table 6
Four-Class Classification Result

Class classification results. We have shown the classification results for Four-Class classification in Table 6.

This team have not actively participated in the competition for the Marathi dataset, but later post-competition implemented all the transformer based models mentioned in this paper and found Macro F1 as 0.8756, which matches with the 3rd rank holder team from the competition.

6. Conclusion

In this shared task, we have compared and evaluated multiple Transformer-based architectures and discovered that XLM-Roberta-large mainly performs better than other Transformer-based models. However, performance varies based on a random seed. It has been observed that by changing random seed, XLM-Roberta performance was impacted less than other Transformer-based models. So, some of the actions will be to identify this observation and speculate the reason behind it. We have also used a couple of Transformer based models like IndicBERT and dehateBERT but was not getting enough raising performance compared with XLMRobeta and mBERT. Our immediate next step will be to investigate the reasons behind the lower performance of IndicBERT and dehateBERT, as IndicBERT is specifically pretrained with Indian

languages and dehateBERT is an already fine-tuned model on the hate speech dataset.

References

[1] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of hate speech in online social media, in: Proceedings of WebSci, ACM, 2019.

[2] C. Newton, The terror queue, 2019. URL: https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video.

[3] A. Robertson, Facebook says ai has fueled a hate speech crackdown, 2020. URL: https://www.theverge.com/2020/11/19/21575139/facebook-moderation-ai-hate-speech.

[4] B. Perrigo, Facebook's hate speech algorithms leave out some languages, 2019. URL: https://time.com/5739688/facebook-hate-speech-languages/.

[5] M. Das, P. Saha, R. Dutt, P. Goyal, A. Mukherjee, B. Mathew, You too brutus! trapping hateful users in social media: Challenges, solutions & insights, 2021, pp. 79–89. doi:10.1145/3465336.3475106.

[6] M. Das, B. Mathew, P. Saha, P. Goyal, A. Mukherjee, Hate speech in online social media, ACM SIGWEB Newsletter (2020) 1–8. doi:10.1145/3427478.3427482.

[7] 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.

[8] 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/.

[9] 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.

[10] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, 2020. arXiv:2012.10289.

[11] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 2012, pp. 71–80. doi:10.1109/SocialCom-PASSAT.2012.55.

[12] P. L. Teh, C.-B. Cheng, W. M. Chee, Identifying and categorising profane words in hate speech, in: Proceedings of the 2nd International Conference on Compute and Data Analysis, ICCDA 2018, Association for Computing Machinery, New York, NY, USA, 2018, p. 65–69. URL: https://doi.org/10.1145/3193077.3193078. doi:10.1145/3193077.3193078.

[13] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and

Trust and 2012 International Confernece on Social Computing, 2012, pp. 71–80. doi:10.1109/SocialCom-PASSAT.2012.55.

[14] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: ICWSM, 2017.

[15] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Detecting offensive language in tweets using deep learning, ArXiv abs/1801.04433 (2018).

[16] Y. Goldberg, A primer on neural network models for natural language processing, Journal of Artificial Intelligence Research 57 (2015). doi:10.1613/jair.4992.

[17] G. L. D. la Peña Sarracén, R. G. Pons, C. E. Muñiz-Cuza, P. Rosso, Hate speech detection using attention-based lstm, in: EVALITA@CLiC-it, 2018.

[18] G. I. Sigurbergsson, L. Derczynski, Offensive language and hate speech detection for Danish, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3498–3508. URL: https://aclanthology.org/2020.lrec-1.430.

[19] Z. Pitenis, M. Zampieri, T. Ranasinghe, Offensive language identification in greek, in: LREC, 2020.

[20] Çagri Çöltekin, A corpus of turkish offensive language on social media, in: LREC, 2020.

[21] 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, Proceedings of the 11th Forum for Information Retrieval Evaluation (2019).

[22] S. Mishra, 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages, in: FIRE, 2019.

[23] S. Renjit, S. M. Idicula, Cusatnlp@hasoc-dravidian-codemix-fire2020:identifying offensive language from manglishtweets, ArXiv abs/2010.08756 (2020).

[24] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, ArXiv abs/1706.03762 (2017).

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019.

[26] 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.

[27] 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, 2020. arXiv:1911.02116.

[28] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, A deep dive into multilingual hate speech classification, in: Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, Springer International Publishing, 2021, pp. 423–439.

[29] S. Mannor, D. Peleg, R. Rubinstein, The cross entropy method for classification, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 561–568. URL: https://doi.org/10.1145/1102351.1102422. doi:10.1145/1102351.1102422.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,

M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. `arXiv:1910.03771`.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, 2019. `arXiv:1912.01703`.

[32] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. `arXiv:1711.05101`.