Transformer Models for Offensive Language Identification in Marathi

Mayuresh Nene[1], Kai North[1], Tharindu Ranasinghe[2] and Marcos Zampieri[1]

[1]Rochester Institute of Technology, USA
[2]University of Wolverhampton, UK

Abstract
This paper describes the WLV-RIT entry to the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) shared task of 2021. The HASOC 2021 organizers provided participants with annotated datasets containing social media posts of English, Hindi and Marathi. We participated in Marathi Subtask 1A: identifying hateful, offensive and profane content. In our methodology, we take advantage of available data from high resource languages by applying cross-lingual transformer-based models and transfer learning to make predictions to Marathi data. Our system achieved a macro F1 score of 0.91 for the test set and it ranked 1st place out of 25 systems.

Keywords
offensive language identification, transformers, Marathi,

1. Introduction

All across the world, millions of smart devices connect to social media platforms on a daily basis, such as Facebook, Twitter, or Instagram. Terabytes of comments, tweets, or posts are uploaded ranging from the user's breakfast to their opinions on global politics. In recent years, there has been a rise in offensive and hateful content [1, 2]. This content is problematic as it may harm the user's mental health [3], and even incite self-harm [4] or violence towards others [5]. It has also be linked to the discrimination or marginalization of particular demographics [6], and to the spread of misinformation [7], causing civil unrest or disobedience [8].

Private entities as well as government bodies are interested in the development of machine learning (ML) models that can automatically identify offensive content on social media [5, 9]. Studies have introduced a variety of hate speech identification systems. These systems have utilized traditional models, such as random forests (RFs) [10, 11], support vector machines (SVMs) [10, 12, 13, 14], and Naive Bayes (NB) [15, 16, 17], to more recent, deep learning [18, 19, 20, 21] and transformer-based models [20, 22, 23, 24]. However, none of these systems are perfect. Issues in dataset quality [2] as well as subjectivity in what is deemed as "offensive content" are still consider major challenges in this task [1, 2]. In addition, the vast majority of hate speech identification systems deal only with English, with exceptions being made for Arabic [25], Greek

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
mn1789@rit.edu (M. Nene); kn1473@g.rit.edu (K. North); T.D.RanasingheHettiarachchige@wlv.ac.uk (T. Ranasinghe); marcos.zampieri@rit.edu (M. Zampieri)

[26], Spanish [27], Hindi [28], and German [29, 30]. Little research has been conducted on under-resourced languages, such as Marathi [31], presenting a gap in the current literature.

In this paper, we present (in detail in Section 4) the WLV-RIT entry to the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) shared-task of 2021. With the aim of providing research on the under-resourced Marathi language, we participated on the Marathi track for sub-task 1A (Section 3.1). We adopted a transfer-learning based approach. We achieved this by experimenting with transformer-based models, such as BERT [32] and XLM-R [33]. These models automatically applied features salient with offensive content in Hindi, to the provided Marathi dataset (Section 4.2). Our system ranked 1st among 25 other systems, having attained a macro F1-score of 0.91.

2. Related Work

Offensive language identification has been a recurrent theme in recent shared tasks. There have been many shared tasks organized in recent years such as OffensEval [34, 35], HASOC [1, 2], TRAC [28, 36], HatEval [37], GermEval [29, 30], and IberEval [27]. Furthermore, there have been different types of offensive content addressed in these shared tasks including hate speech [38], aggression [28, 36], and cyberbullying [39]. In the following section we explain previous editions of HASOC.

2.1. Previous Shared-Tasks in HASOC

Prior to HASOC 2021, two shared-tasks were also organized by the Forum for Information Retrieval Evaluation (FIRE): (1). HASOC 2019 [1], and (2). HASOC 2020 [2]. These shared-tasks challenged participating teams to automatically identify offensive content crawled from Twitter and Facebook. These shared-tasks focused on offensive content in a variety of Indo-European languages, namely English, Hindi, and German. However, HASOC 2020 and HASOC 2021 have since expanded this focus to include several languages with less available data, such as Marathi.

HASOC 2019. HASOC 2019 [1] extracted approximately 7,005 English, 5,983 Hindi, and 4669 German posts from Facebook's and Twitter's APIs. These posts were acquired by searching for hashtags and keywords that were considered offensive by the dataset's authors [1]. The shared-task was split into three sub-tasks: (1). binary classification into hateful (HOF) or non-hateful (NOT) tweets, (2). fine-grained classification into hate speech (HATE), offensive (OFFN), or profanity (PFRN), and (3). recipient identification.

HASOC 2019 provided evidence in favor of deep learning models for hate speech identification. Deep learning models such as a long-short term memory (LSTM) model (YNU) [18], a convolutional neural network model (QutNocturnal) [19], and a BERT model (BRUMS) [23] did exceptionally well, having achieved marco F1 scores for sub-task 1 of 0.7891 for English, 0.8025 for Hindi, and 0.5881 for German respectively [1]. However, questions were raised in regards to the dataset's quality and what affect this may have had on systems' overall performance. It was pointed out that the dataset's reliance on hashtags and keywords provided by the authors, had likely made the dataset prone to the authors' own bias in what they believed to be offensive

or non-offensive content. It was argued that this likely limited the dataset's scope, by not including offensive or controversial topics which were unfamiliar to the authors. In turn, this may have hindered the participating systems' ability to recognize specific forms of hate speech, or offensive content in general.

HASOC 2020. HASOC 2020 [2] attempted to address the validity concerns of HASOC 2019. Instead of using a "hand crafted list of hate speech related terms" [2] to extract offensive posts, the organizer's of HASOC 2020 adopted a randomized sampling technique designed to reduce the impact of the author's bias on dataset quality. An archive containing Tweets in English, Hindi, and German from May 2019 was download from archive.org [2] and used to train a weak binary SVM classifier. 2,600 tweets were identified by the classifier as being hateful. These tweets were copied into HASOC's 2020 new dataset as being examples of hateful tweets. 5% of the remaining 35,000 identified non-hateful tweets were randomly selected and then also added to this dataset. All of the selected tweets were then manually annotated by English, Hindi, and German speaking annotators to produce the dataset's final labels.

HASOC 2020 maintained sub-task 1 and 2 from HASOC 2019 and applied these sub-tasks to its new dataset. Again, deep learning models were found to achieve the best results for sub-task 1. An LSTM model with GloVe word embeddings (IIIT_DWD) [40], a BiLSTM model with fastText word embedings (NSIT) [41], and an ensemble of BERT, DistilBERT, and RoBERTa models (ComMA) [22], attained F1 macro-average scores of 0.86 for English, 0.5337 for Hindi, and 0.5235 for German respectively [2]. Those systems that applied transfer learning also performed well, with cross-lingual models, such as XLM-R, increasing these systems' macro-average scores in some instances [22, 42]. Howbeit, overall performances for this shared-task were considered to be lower than those achieved in HASOC 2019. This was believed to be due to the different sampling technique used, despite it being more realistic.

3. HASOC 2021

3.1. Task Description

HASOC 2021 [43, 44] tasked participating teams to the same sub-tasks of HASOC 2019 [1] and HASOC 2020 [2]. Sub-tasks 1A and 1B being available in English and Hindi, whereas only sub-task 1A being available in English, Hindi, as well as Marathi [44]. Being interested in Marathi, we took part in sub-task 1A.

- **Sub-task 1A:** A binary classification task, whereby participating systems were required to classify tweets into two classes: hateful and offensive (HOF), or non-hateful and non-offensive (NOT);
- **Sub-task 1B:** A more fine-grained classification task, whereby the previously identified HOF posts were further classified into hate speech (HATE), offensive (OFFN), and profanity (PRFN).

An additional sub-task was also made available. Sub-task 2 focused on the use of code-mixed tweets, such as those in Hinglish. Hinglish being a mix of Hindi and English displaying lexemes,

morphology, and syntax taken from both languages. This sub-task also took into consideration the target tweet, referred to as the parent tweet, as well as that tweet's comments, and replies.

- **Sub-task 2:** A binary classification task, whereby participating systems were required to classify code-mixed parent tweets into two classes: hateful and offensive (HOF), or non-hateful and non-offensive (NOT).

3.2. Dataset

The Marathi Offensive Language Dataset (MOLD) [31] is the first dataset of its kind compiled for Marathi. It contains 2,499 tweets extracted from Twitter's API by searching for 22 common Marathi curse words [2, 31]. Non-offensive tweets were obtained by searching for a set of Marathi phrases related "to politics, entertainment, and sports along with the hashtag *#Marathi*" [31]. 6 Marathi speaking annotators then labeled these tweets with the offensive (OFF) or non-offensive (NOT) labels. For HASOC 2021's sub-task 1A, MOLD had a 80%/20% training and test set split.

4. Methods

The methodology applied in this work is divided in two parts. Section 4.1 describes traditional machine learning models that we applied to sub-task 1A, and in Section 4.2 we describe our transformer-based models.

4.1. Traditional Machine Learning Methods

In the first part of the methodology, we used traditional machine learning models. We experimented with three models; Multi-layer Perceptron (MLP) [45], Support Vector Classification (SVC)[46], and RF [47]. The models take an input vector and output a label, either HOF or NOT. The models for MLP, SVC and RF were implemented using *scikit-learn* [48].

Data Preprocessing. Data preprocessing on the traditional 'Devnagri' scripts in which Marathi and many other Indo-Aryan languages are written, included various steps, some of which entailed the removal of stopwords, punctuation marks, URLs, tab spaces. Once we had the most useful data, we went forward with tokenization using the IndicNLP[1] library. We then used TF-IDF to get vectors for the tokens that we had generated. These were then inputted into several traditional machine learning models.

Hyper Parameter Optimization. The SVC model was run with a Grid Search parameter list. This was run on kernel value set to 'rbf', gamma value being selected from '1e-3' and '1e-4' and C value being selected from [1, 10, 100, 1000]. However, the best estimator did not give more than a 1% improvement in the target precision score. The Random Forest Classifier was also run with a grid search parameter list. This was run on 'n_estimators' set between values

[1]The IndicNLP framework is available on https://indicnlp.ai4bharat.org/home/

100, 200, 300, 500 and the 'criterion' being selected between 'gini' and 'entropy'. However, similar to the case with SVC, there was no major improvement in the scores for the target class.

4.2. Transformers

As the second part of the methodology, we used transformer-based models. Transformer architectures have been popular in text classification tasks such as offensive language identification [20, 22, 23]. Their success in these tasks, as seen in HASOC 2020 [2], motivated us to use transformer models for offensive language identification in Marathi.

Pre-trained Transformer models. We experimented with several SOTA transformer models that support Marathi: multilingual BERT (BERT-m) [32] and XLM-Roberta (XLM-R) [33]. XLM-R has an additional advantage: the embeddings are cross-lingual. This helps facilitate transfer learning across languages, as presented later in this section. We followed the same architecture described in Ranasinghe and Zampieri [49] where a simple softmax layer is added to the top of the classification ([CLS]) token to predict the probability of a class label. For XLM-R, from the available two pre-trained models, we specifically used the XLM-R base model. Since the transformer models are prone to random seed [50], each experiment was conducted with three random seeds and considered the majority vote ensemble to get the final result [51].

Transfer Learning. The main appeal of transfer learning is its potential to leverage models trained on data from outside the domain of interest. This can be particularly helpful for boosting the performance of learning on low-resource languages such as Marathi. In this experiments, we used Hindi data released for HASOC 2019 [1]. Hindi is closely related to Marathi, and has high language resources compared to Marathi. Therefore, performing transfer learning form Hindi to Marathi can improve the results of Marathi [52, 53].

We first trained the transformer model separately on HASOC 2019 dataset. Then we saved the weights of the transformer model and the softmax layer and used these weights to initialize the weights of the transformer-based classification model for Marathi.

Implementation. We used a Nvidia Tesla K80 GPU to train the models. We mainly fine tuned the learning rate and number of epochs of the classification model manually to obtain the best results for the validation set. We obtained $1e^{-5}$ as the best value for learning rate and 3 as the best value for number of epochs for all the languages. Training for Hindi language took around 40 minutes while training for Marathi took around 20 minutes. Our implementation is based on HuggingFace [54][2].

5. Results and Evaluation

In this section, we report the experiments we conducted and their results. As informed by the task organizers, we used macro F1 score to measure the model performance. We also report

[2]The implementation is available on https://github.com/tharindudr/DeepOffense

precision, recall and F1 score for each class label as well the macro F1 score in the results tables. The reported results are for the test set.

	NOT			HOF			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
XLM-R (TL)	0.94	0.96	0.94	0.82	0.78	0.79	0.89	0.89	0.91	**0.91**
BERT-m (TL)	0.92	0.94	0.92	0.81	0.77	0.79	0.87	0.87	0.89	0.89
XLM-R	0.90	0.90	0.90	0.79	0.75	0.76	0.86	0.86	0.87	0.88
BERT-m	0.88	0.89	0.89	0.78	0.73	0.74	0.85	0.85	0.86	0.86
Random Forest	0.81	0.87	0.84	0.68	0.58	0.63	0.77	0.77	0.77	0.73
MLP	0.82	0.79	0.80	0.61	0.65	0.63	0.75	0.74	0.75	0.74
SVC	0.79	0.93	0.85	0.78	0.49	0.61	0.79	0.79	0.77	0.79

Table 1
Results for offensive language detection with different machine learning models. For each model, precision (P), recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed. *TL* indicated the *Transfer Learning* experiments. The best result is highlighted in bold.

As can be seen in Table 1, transformer models outperformed the traditional machine learning algorithms. Between the two transformer models, XLM-R performed better than BERT-m. Furthermore, it is clear that transfer learning boosted the results of the transformer models. Our best model was when transfer learning was performed from Hindi with the XLM-R model. According to the results provided by the organisers, our best model scored a 0.91 macro F1 score on the test set and ranked 1^{st} out of 25 participants.

6. Error Analysis

To deepen our understanding of the limitations of our model, we performed an error analysis on the model's results. We compared the predicted labels to the actual labels in multiple models. There were instances where the models falsely predicted the 'HOF' label and instances where they falsely predicted the 'NOT' label. There were some offensive words for which the tweets were almost consistently predicted as non-offensive when the tweets were, in fact, offensive. There are cases where one word has two different meanings. When we use this word with two sets of words, the meaning changes significantly. Two such words are 'गोट्या' (balls) and 'तोंडात' (inside the mouth). While we see that the phrase 'गोट्या तों'ात घेया' means 'eat my balls', the phrase 'गोट्या खेळणे' is essentially a slang for 'not doing a single thing', but with the right context it can also mean 'playing with marbles'. The models have trouble differentiating between such ambiguity where one word can be the same in three very different kinds of statements. False positives for the cases in the figure above have primarily been seen in experiments like the Mulit-Layer Perceptron classifier and the XLM-R model. There are also other instances which the model misses the actual prediction (HOF) and wrongly predicts the text as 'NOT'. The offensive words in the majority of these tweets are relatively lightly offensive as compared to the others. The Random Forest Classifier and the SVC generally falsely predicted these tweets as 'NOT'.

Some other examples of relatively light offensive words where the models have a hit or miss performance are- 'मंद'' which means('dumb or stupid'), 'गद्दार'', which means 'traitor' and

‘पाकिस्तानी’’, which means ‘Pakistani’. The words, such as ‘Pakistani’, would not be termed as offensive without context, which is the tension between the two countries in this case. This leads to a lot of offensive tweets having this word, basically saying ‘Pakistani man’, intended to be an insult at someone who is supposedly talking about harming national security or having said anti-national statements. The models tend to generate a false negative in such cases. It is clear from our learning that the models suffer in cases of ambiguity in meanings and also contextual inference. Hence, such cases need to be handled in a proper way.

7. Conclusion

In this paper we have presented the system submitted by the WLV-RIT team to the HASOC 2021 - Hate Speech and Offensive Content Identification in Marathi at FIRE 2021. We have shown that XLM-R with transfer learning from a closely related language is the most successful transformer model from several transformer models we experimented. We also experimented with a couple of traditional machine learning algorithms. However, the results show that transformer models comfortably outperformed these traditional machine learning models. Our best system, based on XLM-R and transfer learning from Hindi, ranked 1^{st} place out of 25 participants in Marathi.

Acknowledgments

We would like to thank the HASOC organizers for running this interesting shared task and for replying promptly to all our inquiries.

References

[1] 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 FIRE, 2019.

[2] 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: Proceedings of FIRE, 2020.

[3] R. Bannink, S. Broeren, P. M. van de Looij-Jansen, F. G. de Waart, H. Raat, Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents, PloS one 9 (2014).

[4] A. John, A. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, K. Hawton, Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review, Journal of Medical Internet Research 20 (2018).

[5] 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).

[6] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, H. M. H. López, Internet, social media and online hate speech. systematic review, Aggression and Violent Behavior 58 (2021).

[7] N. DePaula, K. J. Fietkiewicz, T. J. Froehlich, A. Million, I. Dorsch, A. Ilhan, Challenges for social media: Misinformation, free speech, civic engagement, and data regulations., in: Proceedings ASIS&T, 2018.
[8] G. De Gregorio, N. Stremlau, Information interventions and social media, Internet Policy Review 10 (2021).
[9] A. Akins, Facebook's oversight board overrules 4 hate speech, misinformation takedowns, SNL Kagan Media and Communications Report (2021).
[10] J. Salminen, H. Almerekhi, M. Milenkovic, S.-g. Jung, A. Jisun, H. Kwak, B. J. Jansen, Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media, in: Proceedings of ICWSM, 2018.
[11] K. Nugroho, E. Noersasongko, Purwanto, Muljono, A. Z. Fanani, Affandy, R. S. Basuki, Improving random forest method to detect hatespeech and offensive word, in: Proceedings of ICOIACT, 2019.
[12] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media, in: Proceedings of NAACL, 2012.
[13] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving cyberbullying detection with user context, in: Proceedings of ECIR, 2013.
[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of WWW, 2016.
[15] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, in: Proceedings of ICWSM, 2011.
[16] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: Proceedings of ASE, 2012.
[17] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: Proceedings of AAAI, 2013.
[18] B. Wang, Y. Ding, S. Liu, X. Zhou, YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language, in: Proceedings of FIRE, 2019.
[19] M. A. Bashar, R. Nayak, QutNocturnal at HASOC 2019: CNN for Hate Speech and Offensive Content Identification in Hindi Language, in: Proceedings of FIRE, 2019.
[20] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hatemonitors: Language agnostic abuse detection in social media, in: Proceedings of FIRE, 2019.
[21] H. Hettiarachchi, T. Ranasinghe, Emoji powered capsule network to detect type and target of offensive posts in social media, in: Proceedings of RANLP, 2019.
[22] R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, ComMA at HASOC 2020: Exploring Multilingual Joint Training across different Classification Tasks, in: Proceedings of FIRE, 2020.
[23] T. Ranasinghe, M. Zampieri, H. Hettiarachchi, BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification, in: Proceedings of FIRE, 2019.
[24] D. Sarkar, M. Zampieri, T. Ranasinghe, A. Ororbia, fBERT: A Neural Transformer for Identifying Offensive Content, in: Proceedings of EMNLP Findings, 2021.
[25] H. Mubarak, D. Kareem, M. Walid, Abusive language detection on Arabic social media, in: Proceedings of ALW, 2017.
[26] Z. Pitenis, M. Zampieri, T. Ranasinghe, Offensive Language Identification in Greek, in:

Proceedings of LREC, 2020.
[27] M. A. Carmona, E. Guzmán-Falcón, M. Montes, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Proceedings of IberEval, 2018.
[28] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking Aggression Identification in Social Media, in: Proceedings of TRAC, 2018.
[29] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the GermEval 2018 shared task on the identification of offensive language, in: Proceedings of GermEval, 2018.
[30] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of germeval task 2, 2019 shared task on the identification of offensive language, in: Proceedings of GermEval, 2019.
[31] 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.
[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL, 2019.
[33] 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, in: Proceedings of ACL, 2019.
[34] 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 SemEval, 2019.
[35] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, c. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), in: Proceedings of SemEval, 2020.
[36] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of TRAC, 2020.
[37] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of SemEval, 2019.
[38] S. Malmasi, M. Zampieri, Challenges in Discriminating Profanity from Hate Speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 1 – 16.
[39] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, Computers in Human Behavior 93 (2019) 333–345.
[40] A. Mishra, S. Saumya, A. Kumar, IIITDWD at HASOC 2020: Identifying offensive content in multitask Indo-European languages, in: Proceedings of FIRE, 2020.
[41] R. Raj, S. Srivastava, S. Saumya, NSIT and IIITDWD at HASOC 2020: Deep learning model for hate-speech identification in Indo-European languages, in: Proceedings of FIRE, 2020.
[42] X. Ou, H. Li, YNU_OXZ at HASOC 2020: Multilingual Hate Speech and Offensive Content Identification based on XLM-RoBERTa, in: Proceedings of FIRE, 2020.
[43] Modha, Sandip and Mandl, Thomas and Shahi, Gautam Kishore and Madhu, Hiren and Satapara, Shrey and Ranasinghe, Tharindu and Zampieri, Marcos, 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.
[44] 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.
[45] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (1991) 183–197.
[46] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.
[47] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (2002) 18–22.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.
[49] T. Ranasinghe, M. Zampieri, Multilingual Offensive Language Identification with Cross-lingual Embeddings, in: Proceedings of EMNLP, 2020.
[50] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, Revisiting few-sample bert fine-tuning, in: Proceedings of ICLR, 2020.
[51] H. Hettiarachchi, T. Ranasinghe, Infominer at wnut-2020 task 2: Transformer-based covid-19 informative tweet extraction, in: Proceedings of W-NUT, 2020.
[52] T. Ranasinghe, M. Zampieri, Multilingual Offensive Language Identification for Low-resource Languages, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) (2021).
[53] T. Ranasinghe, M. Zampieri, An evaluation of multilingual offensive language identification methods for the languages of india, Information 12 (2021).
[54] 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. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of EMNLP, 2020.