Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi

Saurabh Gaikwad[1], Tharindu Ranasinghe[2], Marcos Zampieri[1], Christopher M. Homan[1]
[1]Rochester Institute of Technology, USA
[2]University of Wolverhampton, UK
`T.D.RanasingheHettiarachchige@wlv.ac.uk`

Abstract

The widespread presence of offensive language on social media motivated the development of systems capable of recognizing such content automatically. Apart from a few notable exceptions, most research on automatic offensive language identification has dealt with English. To address this shortcoming, we introduce *MOLD*[1], the Marathi Offensive Language Dataset. *MOLD* is the first dataset of its kind compiled for Marathi, thus opening a new domain for research in low-resource Indo-Aryan languages. We present results from several machine learning experiments on this dataset, including zero-short and other transfer learning experiments on state-of-the-art cross-lingual transformers from existing data in Bengali, English, and Hindi.

1 Introduction

The presence of hate speech, cyber-bullying, and other forms of offensive language in online communities is a global phenomenon. Even though thousands of languages and dialects are widely used in social media, most studies on the automatic identification of such content consider English only, a language for which datasets and other resources such as pre-trained models exist (Rosenthal et al., 2021). In the past few years researchers have studied this problem on languages such as Arabic (Mubarak et al., 2021), French (Chiril et al., 2019), and Turkish (Çöltekin, 2020) to name a few. In doing so, they have created new datasets for each of these languages. Competitions such as OffensEval (Zampieri et al., 2020) and TRAC (Kumar et al., 2020a) provided multilingual datasets, which enabled the use of data augmentation methods (Ghadery and Moens, 2020), multilingual word embeddings (Pamungkas and Patti, 2019), and cross-

lingual contextual word embeddings (Ranasinghe and Zampieri, 2020) to tackle this problem.

In this paper, we revisit the task of offensive language identification for low resource languages, focusing on Marathi, an Indo-Aryan language spoken by over 80 million people, most of whom live in India. Even though Marathi is spoken by a large population, it is relatively low-resourced compared to other languages spoken in the region. We collect and annotate the first Marathi offensive language identification dataset to date and we train a number of monolingual models on this dataset. Finally, we explore state-of-the-art cross-lingual learning methods to project predictions to Marathi from Bengali, Hindi, and English. We address two research questions in this paper:

RQ1: What is the impact of the dataset size in monolingual and cross-lingual models for offensive language identification? While the Marathi dataset is relatively small, cross-lingual transfer learning methods allow us to take advantage of larger available datasets in other languages.

RQ2: What is the influence of language similarity in cross-lingual predictions for offensive language identification? Previous work used English as the base language to make predictions in lower resourced languages. In this paper we use two Indo-Aryan languages, Bengali and Hindi, to project predictions into Marathi.

Our main contributions are the following:

1. We release *MOLD*, the Marathi Offensive Language Dataset, with nearly 2,500 annotated tweets. *MOLD* is the first dataset for offensive language identification in Marathi.

2. We evaluate the performance of several traditional machine learning models (e.g. SVMs)

---

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

and deep learning models (e.g. LSTM) trained on *MOLD*.

3. We apply cross-lingual transformers to offensive language identification in Marathi. We take advantage of existing data in English and in two Indo-Aryan languages, Hindi and Bengali, to project predictions to Marathi and we compare the results of these strategies. To the best of our knowledge, this is the first paper to study closely-related languages in transfer learning for offensive language identification.

4. In addition to *MOLD*, we make the code and the models freely available to the community.

2 Related Work

The problem of offensive content online has been widely studied using computational models. Researchers have trained system to recognize various types of such content such as *cyberbulling*, *hate speech*, and many others. In terms of computational approaches, early studies approached the problem using feature engineering and classical machine learning classifiers, most notably SVMs (Dadvar et al., 2013; Malmasi and Zampieri, 2017), while more recent work applied deep neural networks combined with word embeddings (Aroyehun and Gelbukh, 2018; Hettiarachchi and Ranasinghe, 2019). With the development of large pre-trained transformer models such as BERT and XLNET (Devlin et al., 2019; Yang et al., 2019), several studies have explored the use of general pre-trained transformers in offensive language identification (Liu et al., 2019; Ranasinghe et al., 2019; Bucur et al., 2021) as well retrained or fine-tuned models on offensive language corpora such as HateBERT (Caselli et al., 2020).

While the vast majority of studies address offensive language identification using English data (Yao et al., 2019; Ridenhour et al., 2020), several recent studies have created new datasets for various languages and applied computational models to identify such content in Arabic (Mubarak et al., 2021), Dutch (Tulkens et al., 2016), French (Chiril et al., 2019), German (Wiegand et al., 2018), Greek (Pitenis et al., 2020), Hindi (Bohra et al., 2018), Italian (Poletto et al., 2017), Portuguese (Fortuna et al., 2019), Slovene (Fišer et al., 2017), Spanish (Plaza-del Arco et al., 2021), and Turkish (Çöltekin, 2020). A recent trend is the use of pre-trained multilingual models such as XLM-R (Conneau et al., 2019) to leverage available English resources to make predictions in languages with less resources (Plaza-del Arco et al., 2021; Ranasinghe and Zampieri, 2020, 2021c,b; Sai and Sharma, 2021). This is made possible by the availability of the aforementioned datasets as well multilingual datasets made available at shared tasks such as HASOC 2019 (Mandl et al., 2019), TRAC 2018 and 2020 (Kumar et al., 2018, 2020a), and two tasks at SemEval: HatEval 2018 (Basile et al., 2019) and OffensEval 2020 (Zampieri et al., 2020).

3 Datasets

We present *MOLD* and four other datasets used in this work: the Bengali dataset (Bhattacharya et al., 2020) used in the TRAC-2 shared task (Kumar et al., 2020a)—henceforth *BE*, the Hindi dataset (Mandl et al., 2019) used in the HASOC 2019 shared task—henceforth *HI*, and the English datasets used in OffensEval, SemEval-2019 Task 6 and SemEval-2020 Task 12—henceforth *EN-OLID* (Zampieri et al., 2019) and *EN-SOLID* (Rosenthal et al., 2021), respectively.

To annotate *MOLD*, we followed OLID's annotation scheme for English which has been replicated in SOLID and in datasets in Greek (Pitenis et al., 2020), Turkish (Çöltekin, 2020) and many other languages. OLID's taxonomy comprises the following three levels:

Level A: Offensive language identification: offensive (OFF) vs. non-offensive (NOT)

Level B: Categorization of offensive language: targeted insult or thread vs. untargeted profanity.

Level C: Offensive language target identification: individual vs. group vs. other.

This hierarchical taxonomy represents multiple types of offensive content in a single annotation scheme (e.g. targeted insults to an individual are often *cyberbullying* and targeted insults to a group are often *hate speech*) making it a great fit for cross-lingual learning applied to low-resource languages like Marathi. We used OLID level A labels to annotate *MOLD* and we map these labels to those included in the Bengali and Hindi datasets.

MOLD The Marathi dataset contains data collected from Twitter using the Twitter API. We aimed to achieve a similar distribution of offensive vs. non-offensive content present in OLID, which contains around 33% offensive and 67% non-offensive tweets. To make sure that both classes

were represented, we used both offensive and non-offensive keywords. For the offensive content we used 22 common curse words in Marathi and for the non-offensive content we used search phrases related to politics, entertainment, and sports along with the hashtag #Marathi.

We collected a total 2,547 tweets that were annotated by 6 volunteer annotators who are native speakers of Marathi with age between 20 and 25 years old and a bachelors degree. The annotation task is a binary classification, in which annotators assigned tweets as offensive (OFF) or not offensive (NOT). The annotators could flag a tweet as invalid if it contained four or more non-Marathi words. The final version of *MOLD* contains 2,499 annotated tweets randomly split 75%/25% into training and testing sets, respectively. We used Cohen's kappa (Carletta, 1996) to measure agreement between pairs of annotators. We provided a common set of 100 instances to each of the three pairs of annotators and we report scores of 0.91 between A1 and A2, 0.79 between A3 and A4, and 0.77 between A5 and A6. Table 1 shows dataset statistics, including class distribution.

Class	Training	Testing	Total
Not Offensive	1,205	418	1,623
Offensive	669	207	876
Total	1,874	625	2,499

Table 1: Number of instances and class distribution of NOT and OFF tweets in *MOLD*.

Other Datasets In addition to the Marathi dataset, we used the four aforementioned publicly available offensive language detection datasets presented in Table 2. OLID (*EN-OLID*) is one of the most popular offensive language datasets for English and we used its level A annotations (offensive vs. non-offensive) as labels. We used *EN-SOLID*, the largest available dataset of its kind as our second English dataset. *EN-SOLID* contains over nine million English tweets labeled in a weakly supervised manner (Rosenthal et al., 2021). *EN-SOLID* was created using an ensemble of four different models and provides, along with the class labels, the average and standard deviation of the confidence scores predicted by each model. We included only training examples with average confidence scores greater than 0.85 over all models, leaving us with 120,758 examples. Using both *EN-OLID* and *EN-SOLID* allows us to investigate the impact of training data size and help us answer **RQ1**.

To perform transfer learning from a closely-related language to Marathi, we used *HI* (Mandl et al., 2019). Both the English and Hindi datasets contain Twitter data making them in-domain with respect to *MOLD*. *BE*, the Bengali dataset (Bhattacharya et al., 2020), is different than the other datasets as it contain Facebook data and three classes, allowing us to compare the performance of cross-lingual embeddings on off-domain data but in a a language similar to Marathi. For Bengali we merged the classes overtly aggressive and covertly aggressive and map them to *EN-OLID*'s offensive class. Using both *BE* and *HI* in addition to the two English datasets allow us to investigate the impact of language similarity aiming to answer our **RQ2**.

4 Methods and Results

4.1 Monolingual Models

We run several computational models on *MOLD*. We trained four classical machine learning classifiers, available in Scikit-learn (Pedregosa et al., 2011): Decision Trees, Naive Bayes, Random Forest, and SVM using bag of words (BoW), word unigrams, and word unigrams and bigrams combines using TF-IDF weighting. We took several pre-processing steps before extracting features such as removing numbers, extra spaces, special characters, and stop words.[2]

We implemented several deep learning models, such as multi layer perceptron (MLP), long short-term memory networks (LSTMs) with embedding layers, and bi-LSTMs with attention and word embedding layers. We used the Marathi word2vec embeddings released in Kumar et al. (2020b). We also experimented with several SOTA transformer models that support Marathi: multilingual BERT (BERT-m) (Devlin et al., 2019) and XLM-Roberta (XLM-R) (Conneau et al., 2019). 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 (2020) 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 large model.

---

[2]Marathi stop words are available on `https://github.com/stopwords-iso/stopwords-mr/blob/master/stopwords-mr.txt`

Code	Language	Dataset	Instances	Source	Labels
BE	Bengali	TRAC	4,000	F	overtly aggressive, covertly aggressive, non aggressive
EN-OLID	English	OLID	14,100	T	offensive, non-offensive
EN-SOLID	English	SOLID	120,758	T	offensive, non-offensive
HI	Hindi	HASOC	8,000	T	hate offensive, non hate-offensive

Table 2: Instances, sources, and labels in all datasets. F stands for Facebook and T for Twitter.

For both classical and deep learning models we finetuned hyperparameters manually to obtain the best results for the validation set created using a 0.8:0.2 split on the training data. As the deep learning models tend to overfit, we evaluated the model on the validation set once in every 100 training batches. We performed *early stopping* if the validation loss did not improve over 10 evaluation steps. All the deep learning experiments were run on an Nvidia Tesla K80 GPU.
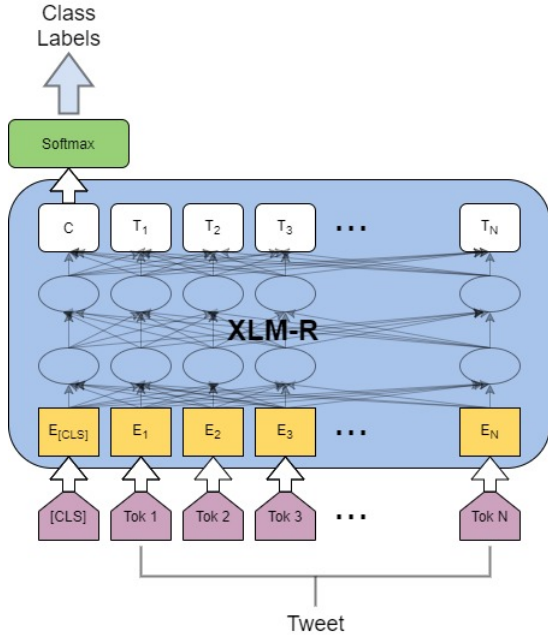

Figure 1: Text classification architecture with XLM-R (Ranasinghe and Zampieri, 2020).

Table 3 shows the results obtained by all monolingual models on *MOLD*'s test set in terms of both Macro F1 and Weighted F1. We use both metrics due to the data imbalance in *MOLD*. With the exception of MLP, all of the deep learning models outperformed the classical ones. This is somewhat surprising as classical models tend to outperform deep models on relatively small datasets like *MOLD*, but it corroborates the findings from recent competitions on this topic (Basile et al., 2019). Of the deep learning models, XLM-R transformers provided the best results with a 0.91 macro F1 score.

Features	Model	M F1	W F1
Embeddings	XLM-R	0.9103	0.9210
Embeddings	BERT-m	0.8852	0.8994
Embeddings	LSTM	0.8400	0.84.09
Embeddings	Bi-LSTM	0.8238	0.8251
BoW	Random Forest	0.7686	0.7796
Embeddings	MLP	0.7541	0.7830
BoW	SVM	0.7489	0.7813
BoW	Naive Bayes	0.7223	0.7597
BoW	Decision Tree	0.7028	0.7395

Table 3: Monolingual results for Marathi ordered by macro (M) F1. We also report weighted (W) F1 scores.

4.2 Cross-lingual Models

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 like Marathi. The recent success of XLM-R cross-lingual transformers with transfer learning in offensive language identification for low resource languages (Ranasinghe and Zampieri, 2020) confirms that this is a feasible approach. In these experiments, however, the transfer learning's base language was English whereas here we use two languages related to Marathi: Bengali and Hindi, in order to evaluate the extent to which language similarity boosts transfer learning performance.

Transfer Learning We first trained the XLM-R model separately on the *BE*, *HI*, *EN-OLID* and *EN-SOLID* datasets. 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. TL row in Table 4 shows the results obtained by the cross lingual models with XLM-R. The use of transfer learning substantially improved the monolingual results. With 8,000 and 4,000 training instances, respectively, the transfer learning model achieved macro F1 scores of 0.9401 from Hindi and of 0.9345 from Bengali, respec-

tively, outperforming the results obtained using the two English datasets, *EN-OLID* and, especially, *EN-SOLID*, each contain more instances than either the Hindi or the Bengali dataset, yet they fail to outperform either as the base dataset in our transfer learning experiments, suggesting that language similarity played a positive role in transfer learning.

Zero shot learning   To further observe the impact of language similarity in transfer learning, we performed Zero shot learning, where the XLM-R model was trained on the other datasets and tested on the Marathi test set. According to the results in Zero-shot row of Table 4 *HI* outperforms all the other languages in Zero shot too.

Methodology	Dataset	M F1	W F1
Transfer Learning	HI	0.9401	0.9492
	BE	0.9345	0.9422
	EN-SOLID	0.9321	0.9399
	EN-OLID	0.9298	0.9385
Zero-Shot	HI	0.8396	0.8461
	BE	0.8115	0.8176
	EN-SOLID	0.7954	0.8004
	EN-OLID	0.7854	0.7901

Table 4: Transfer learning results ordered by macro (M) F1 for Marathi. We also report weighted (W) F1 scores.

Few shot learning   Finally, we evaluated each of the languages performance in few shot learning with Marathi. We retrained offensive language identification XLM-R models from other languages on 100, 200, 300 etc. instances from Marathi. As shown in Figure 2 *HI* tops other languages in all the few shot experiments making it further clear that transfer learning from a more similar language is effective in offensive language identification.

5   Conclusion and Future Work

This paper introduced *MOLD*, the first offensive language dataset for Marathi. We evaluated the performance of several machine learning models trained to identify offensive content in Marathi. Our results show that applying cross-lingual contextual word embeddings substantially improved performance over monolingual models. Furthermore, we showed that XLM-R with transfer learning from Hindi outperforms all of the other methods we tested. The results obtained by our models confirm that closely related languages provide an advantage in our transfer learning experiments, answering our **RQ2**. This is likely due to the fact that
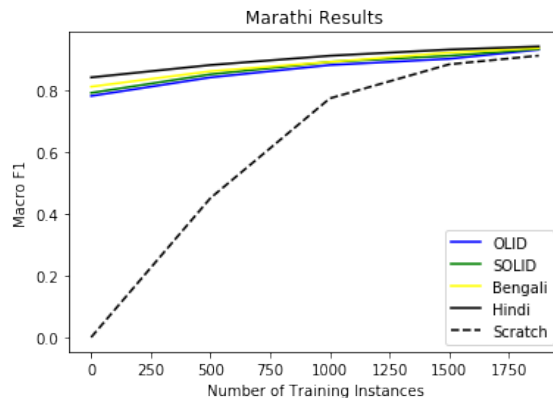

Figure 2: Macro F1 with different number of examples and with different transfer learning strategies for Marathi

Hindi and Marathi are typologically related and also because these languages are in a situation of language contact sharing cultural background.

To the best of our knowledge, this paper is the first to address the question of language similarity in cross-lingual learning for offensive language identification. With respect to our **RQ1**, our results show that the difference in performance between transfer learning strategies from OLID and from SOLID is minimal. SOLID is more than eight times larger than OLID, suggesting that beyond a certain point, more instances do not necessarily yield significant performance improvements in transfer learning. Finally, we believe that the findings presented in this paper can open a wide range of avenues to offensive language identification applied to other low resource languages, particularly from the Indo-Aryan family.

MOLD is the official dataset for Marathi at the HASOC 2021[3] shared task on Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. We are expanding the annotation of this dataset to the levels B and C of OLID's annotation taxonomy. This will provide us with the opportunity to test computational models to identify the type and target of offensive posts in Marathi. As future work, we would like to evaluate the performance of transfer learning from Dravidian languages spoken in India such as Tamil and Telugu to analyze the interplay between language similarity and cultural overlap in cross-lingual offensive language identification as in Ranasinghe and Zampieri (2021a).

---

[3] https://hasocfire.github.io/hasoc/2021/index.html

Acknowledgment

References

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of TRAC*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of TRAC*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of PEOPLES*.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the ACL*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Proceedings of TALN*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Dyberbullying Detection with User Context. In *Proceedings of ECIR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings ALW*.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A Hierarchically-labeled Portuguese Hate Speech Dataset. In *Proceedings of ALW*.

Erfan Ghadery and Marie-Francine Moens. 2020. LIIR at SemEval-2020 task 12: A cross-lingual augmentation approach for multilingual offensive language identification. In *Proceedings of SemEval*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2019. Emoji powered capsule network to detect type and target of offensive posts in social media. In *Proceedings of RANLP*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020a. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Saurav Kumar, Saunack Kumar, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020b. "a passage to India": Pre-trained word embeddings for Indian languages. In *Proceedings of SLTU*.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of SemEval*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of RANLP*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of FIRE*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic Offensive Language on Twitter: Analysis and Experiments. In *Proceedings of WANLP*.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings ACL:SRW*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of LREC*.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of CLiC-it*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021a. An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India. *Information*, 12(8):306.

Tharindu Ranasinghe and Marcos Zampieri. 2021b. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Tharindu Ranasinghe and Marcos Zampieri. 2021c. Multilingual Offensive Language Identification for Low-resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In *Proceedings of FIRE*.

Michael Ridenhour, Arunkumar Bagavathi, Elaheh Raisi, and Siddharth Krishnan. 2020. Detecting online hate speech: Approaches using weak supervision and network embedding models. In *Proceedings of SBP-BRiMS*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *Findings of the ACL*.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for Dravidian languages. In *Proceedings of DravidianLangTech*.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A Dictionary-based Approach to Racism Detection in Dutch Social Media. In *Proceedings of TA-COS*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of NeurIPS*.

Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *Proceedings of WWW*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.