Multilingual Offensive Language Identification
with Cross-lingual Embeddings

Tharindu Ranasinghe
University of Wolverhampton
Wolverhampton, UK
`tharindu.ranasinghe@wlv.ac.uk`

Marcos Zampieri
Rochester Institute of Technology
Rochester, NY, USA
`mazgla@rit.edu`

Abstract

Offensive content is pervasive in social media and a reason for concern to companies and government organizations. Several studies have been recently published investigating methods to detect the various forms of such content (e.g. hate speech, cyberbulling, and cyberaggression). The clear majority of these studies deal with English partially because most annotated datasets available contain English data. In this paper, we take advantage of English data available by applying cross-lingual contextual word embeddings and transfer learning to make predictions in languages with less resources. We project predictions on comparable data in Bengali, Hindi, and Spanish and we report results of 0.8415 F1 macro for Bengali, 0.8568 F1 macro for Hindi, and 0.7513 F1 macro for Spanish. Finally, we show that our approach compares favorably to the best systems submitted to recent shared tasks on these three languages, confirming the robustness of cross-lingual contextual embeddings and transfer learning for this task.

1 Introduction

Offensive posts on social media result in a number of undesired consequences to users. They have been investigated as triggers of suicide attempts and ideation, and mental health problems (Bonanno and Hymel, 2013; Bannink et al., 2014). One of the most common ways to cope with offensive content online is training systems to be capable of recognizing offensive messages or posts. Once recognized, such offensive content can be set aside for human moderation or deleted from the respective platform (e.g. Facebook, Twitter), preventing harm to users and controlling the spread of abusive behavior in social media.

There have been several recent studies published on automatically identifying various kinds of offensive content such as abuse (Mubarak et al., 2017), aggression (Kumar et al., 2018, 2020), cyber-bullying (Rosa et al., 2019), and hate speech (Malmasi and Zampieri, 2018). While there are a few studies published on languages such as Arabic (Mubarak et al., 2020) and Greek (Pitenis et al., 2020), most studies and datasets created so far include English data. Data augmentation (Ghadery and Moens, 2020) and multilingual word embeddings (Pamungkas and Patti, 2019) have been applied to take advantage of existing English resources to improve the performance in systems dealing with languages other than English. To the best of our knowledge, however, state-of-the-art cross-lingual contextual embeddings such as XLM-R (Conneau et al., 2019) have not yet been applied to offensive language identification. To address this gap, we evaluate the performance of cross-lingual contextual embeddings and transfer learning (TL) methods in projecting predictions from English to other languages. We show that our methods compare favorably to state-of-the-art approaches submitted to recent shared tasks on all datasets. The main contributions of this paper are the following:

1. We apply cross-lingual contextual word embeddings to offensive language identification. We take advantage of existing English data to project predictions in three other languages: Bengali, Hindi, and Spanish.

2. We tackle both off-domain and off-task data for Bengali. We show that not only can these methods project predictions for different languages but also for different domains (e.g. Twitter vs. Facebook) and tasks (e.g. binary vs. three-way classification).

3. We provide important resources to the community: the code, and the English model will be freely available to everyone interested in working on low-resource languages using the same methodology.

2 Related Work

There is a growing interest in the development of computational models to identify offensive content online. Early approaches relied heavily on feature engineering combined with traditional machine learning classifiers such as naive bayes and support vector machines (Xu et al., 2012; Dadvar et al., 2013). More recently, neural networks such as LSTMs, bidirectional LSTMs, and GRUs combined with word embeddings have proved to outperform traditional machine learning methods in this task (Aroyehun and Gelbukh, 2018; Majumder et al., 2018). In the last couple of years, transformer models like ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019) have been applied to offensive language identification achieving competitive scores and topping the leaderboards in recent shared tasks (Liu et al., 2019; Ranasinghe et al., 2019). Most of these approaches use existing pretrained transformer models which can also be used as text classification models.

The clear majority of studies on this topic deal with English (Malmasi and Zampieri, 2017; Yao et al., 2019; Ridenhour et al., 2020) partially motivated by the availability English resources (e.g. corpora, lexicon, and pre-trained models). In recent years, a number of studies have been published on other languages such as Arabic (Mubarak et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), Dutch (Tulkens et al., 2016), French (Chiril et al., 2019), Greek (Pitenis et al., 2020), Italian (Poletto et al., 2017), Portuguese (Fortuna et al., 2019), Slovene (Fišer et al., 2017), and Turkish (Çöltekin, 2020) creating new datasets and resources for these languages.

Recent competitions organized in 2020 such as TRAC (Kumar et al., 2020) and OffensEval (Zampieri et al., 2020) have included datasets in multiple languages providing participants with the opportunity to explore cross-lingual learning models opening exciting new avenues for research on languages other than English and, in particular, on low-resource languages. The aforementioned deep learning methods require large annotated datasets to perform well which is not always available for low-resource languages. In this paper, we address the problem of data scarcity in offensive language identification by using transfer learning and cross-lingual transformers from a resource rich language like English to three other languages: Bengali, Hindi, and Spanish.

3 Data

We acquired datasets in English and three other languages: Bengali, Hindi, and Spanish (listed in Table 1). The four datasets have been used in shared tasks in 2019 and 2020 allowing us to compare the performance of our methods to other approaches. As our English dataset, we chose the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a), used in the SemEval-2019 Task 6 (OffensEval) (Zampieri et al., 2019b). OLID is arguably one of the most popular offensive language datasets. It contains manually annotated tweets with the following three-level taxonomy and labels:

A: Offensive language identification - offensive vs. non-offensive;

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

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

We chose OLID due to the flexibility provided by its hierarchical annotation model that considers multiple types of offensive content in a single taxonomy (e.g. targeted insults to a group are often *hate speech* whereas targeted insults to an individual are often *cyberbulling*). This allows us to map OLID level A (offensive vs. non-offensive) to labels in the other three datasets. OLID's annotation model is intended to serve as a general-purpose model for multiple *abusive language detection subtasks* (Waseem et al., 2017). The transfer learning strategy used in this paper provides us with an interesting opportunity to evaluate how closely the OLID labels relate to the classes in datasets annotated using different guidelines and sub-task definitions (e.g. *aggression* and *hate speech*).

The Hindi dataset (Mandl et al., 2019) was used in the HASOC 2019 shared task, while the Spanish dataset (Basile et al., 2019) was used in SemEval-2019 Task 5 (HatEval). They both contain Twitter data and two labels. The Bengali dataset (Bhattacharya et al., 2020) was used in the TRAC-2 shared task (Kumar et al., 2020) on aggression identification. It is different than the other three datasets in terms of domain (Facebook instead of Twitter) and set of labels (three classes instead of binary), allowing us to compare the performance of cross-lingual embeddings on off-domain data and off-task data.

Lang.	Inst.	S	Labels
Bengali	4,000	F	overtly aggressive, covertly aggressive, non aggressive
English	14,100	T	offensive, non-offensive
Hindi	8,000	T	hate offensive, non hate-offensive
Spanish	6,600	T	hateful, non-hateful

Table 1: Instances (Inst.), source (S) and labels in all datasets. F stands for Facebook and T for Twitter.

4 Methodology

Transformer models have been used successfully for various NLP tasks (Devlin et al., 2019). Most of the tasks were focused on English language due to the fact the most of the pre-trained transformer models were trained on English data. Even though, there were several multilingual models like BERT-m (Devlin et al., 2019) there was much speculations about its ability to represent all the languages (Pires et al., 2019) and although BERT-m model showed some cross-lingual characteristics it has not been trained on crosslingual data (Karthikeyan et al., 2020). The motivation behind this methodology was the recently released cross-lingual transformer models - XLM-R (Conneau et al., 2019) which has been trained on 104 languages. The interesting fact about XLM-R is that it is very compatible in monolingual benchmarks while achieving the best results in cross-lingual benchmarks at the same time (Conneau et al., 2019). The main idea of the methodology is that we train a classification model on a resource rich, typically English, using a cross-lingual transformer model, save the weights of the model and when we initialise the training process for a lower resource language, start with the saved weights from English. This process is also known as transfer learning and is illustrated in Figure 1.
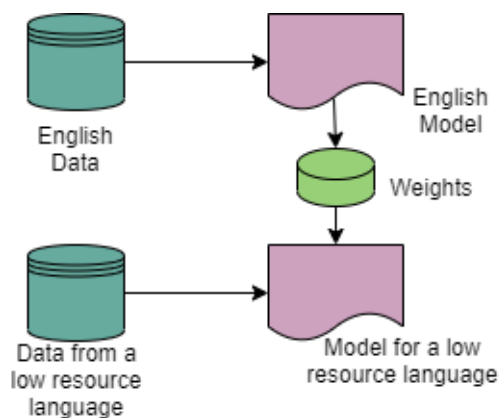

Figure 1: Transfer learning strategy.

There are two main parts of the methodology. Subsection 4.1 describes the classification architecture we used for all the languages. In Subsection 4.2 we describe the transfer learning strategies used to take advantage of English offensive language data in predicting offense in less-resourced languages.

4.1 XLM-R for Text Classification

Similar to other transformer architectures XLM-R transformer architecture can also be used for text classification tasks (Conneau et al., 2019). XLM-R-large model contains approximately 125M parameters with 12-layers, 768 hidden-states, 3072 feed-forward hidden-states and 8-heads (Conneau et al., 2019). It takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The first token of the sequence is always [CLS] which contains the special classification embedding (Sun et al., 2019).

For text classification tasks, XLM-R takes the final hidden state $\mathbf{h}$ of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of XLM-R to predict the probability of label c: as shown in Equation 1 where W is the task-specific parameter matrix.

$$p(c|\mathbf{h}) = softmax(W\mathbf{h}) \tag{1}$$

We fine-tune all the parameters from XLM-R as well as W jointly by maximising the log-probability of the correct label. The architecture diagram of the classification is shown in Figure 2. We specifically used the XLM-R large model.
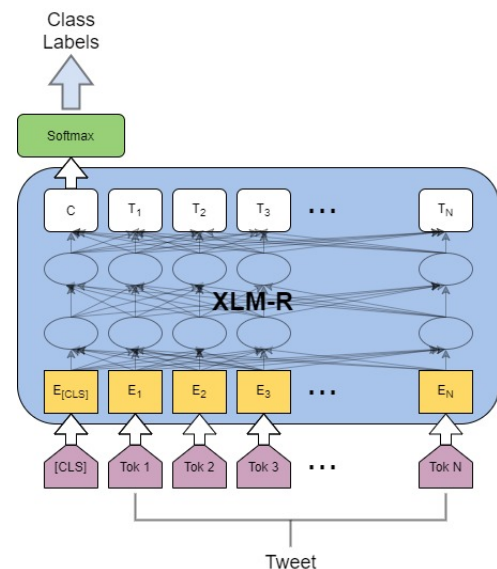

Figure 2: Text classification architecture with XLM-R.

4.2 Transfer-learning strategies

When we adopt XLM-R for multilingual offensive language identification, we perform transfer learning in two different ways.

Inter-language transfer learning We first trained the XLM-R classification model on first level of English offensive language identification dataset (OLID) (Zampieri et al., 2019a). Then we save the weights of the XLM-R model as well as the softmax layer. We use these saved weights from English to initialise the weights for a new language. To explore this transfer learning aspect we experimented on Hindi language which was released for HASOC 2019 shared task (Mandl et al., 2019) and on Spanish data released for Hateval 2019 (Basile et al., 2019).

Inter-task and inter-language transfer learning Similar to the inter-language transfer learning strategy, we first trained the XLM-R classification model on the first level of English offensive language identification dataset (OLID) (Zampieri et al., 2019a). Then we only save the weights of the XLM-R model and use the saved weights to initialise the weights for a new language. We did not use the weights of the last softmax layer since we wanted to test this strategy on data that has a different number of offensive classes to predict. We explored this transfer learning aspect with Bengali dataset released with TRAC - 2 shared task (Kumar et al., 2020). As described in the Section 3 the classifier should make a 3-way classification in between 'Overtly Aggressive', 'Covertly Aggressive' and 'Non Aggressive' text data.

We used a Nvidia Tesla K80 GPU to train the models. We divided the dataset into a training set and a validation set using 0.8:0.2 split on the dataset. We predominantly 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 English language took around 1 hour while training for other languages took around 30 minutes. The code and the pretrained English model is available on GitHub.[1]

5 Results and Evaluation

We evaluate the results obtained by all models using the test sets provided by the organizers of each competition. We compared our results to the best systems in TRAC-2 for Bengali, HASOC for Hindi, HatEval for Spanish in terms of weighted and macro F1 score according to the metrics reported by the task organizers - TRAC-2 reported only macro F1, HatEval reported only weighted F1, and HASOC reported both.

Finally, we evaluate the improvement of the transfer learning strategy in the performance of both BERT and XLM-R. We present the results along with the majority class baseline for each language in Table 2. *TL* indicates that the model used the inter language transfer learning strategy described in Subsection 4.2.

Language	Model	M F1	W F1
Bengali	XLM-R (TL)	0.8415	0.8423
	Risch and Krestel (2020)	0.8219	
	BERT-m (TL)	0.8197	0.8231
	XLM-R	0.8142	0.8188
	BERT-m	0.8132	0.8157
	Baseline	0.2498	0.4491
Hindi	XLM-R (TL)	0.8568	0.8580
	BERT-m (TL)	0.8211	0.8220
	Bashar and Nayak (2019)	0.8149	0.8202
	XLM-R	0.8061	0.8072
	BERT-m	0.8025	0.8030
	Baseline	0.3510	0.3798
Spanish	XLM-R (TL)	0.7513	0.7591
	BERT-m (TL)	0.7319	0.7385
	Vega et al. (2019)		0.7300
	Pérez and Luque (2019)		0.7300
	XLM-R	0.7224	0.7265
	BERT-m	0.7215	0.7234
	Baseline	0.3700	0.4348

Table 2: Results ordered by macro (M) F1 for Bengali and weighted (W) F1 for Hindi and Spanish.

For Hindi, transfer learning with XLM-R cross lingual embeddings provided the best results achieving 0.8568 and 0.8580 weighted and macro F1 score respectively. In HASOC 2019 (Mandl et al., 2019), the best model by Bashar and Nayak (2019) scored 0.8149 Macro F1 and 0.8202 Weighted F1 using convolutional neural networks.

For Spanish transfer learning with XLM-R cross lingual embeddings also provided the best results achieving 0.7513 and 0.7591 macro and weighted F1 score respectively. The best two models in HatEval (Basile et al., 2019) for Spanish scored 0.7300 macro F1 score. Both models applied SVM classi-

fiers trained on a variety of features like character and word n-grams, POS tags, offensive word lexicons, and embeddings.

The results for Bengali deserve special attention because the Bengali data is off-domain with respect to the English data (Facebook instead of Twitter), and it contains three labels (covertly aggressive, overtly aggressive, and not aggressive) instead of the two labels present in the English dataset (offensive and non-offensive). *TL* indicates that the model used the inter-task, inter-domain, and inter-language transfer learning strategy described in Subsection 4.2. Similar to the Hindi and Spanish, transfer learning with XLM-R cross lingual embeddings provided the best results for Bengali achieving 0.7513 and 0.7591 macro and weighted F1 respectively thus outperforming the other models by a significant margin. The best model in the TRAC-2 shared task (Kumar et al., 2020) scored 0.821 weighted F1 score in Bengali using a BERT-based system.

We look closer at the test set predictions by XLM-R (TL) for Bengali in Figure 3. We observe that the performance for the non-aggressive class is substantially better than the performance for the overtly aggressive and covertly aggressive classes following a trend observed by the TRAC-2 participants including Risch and Krestel (2020).
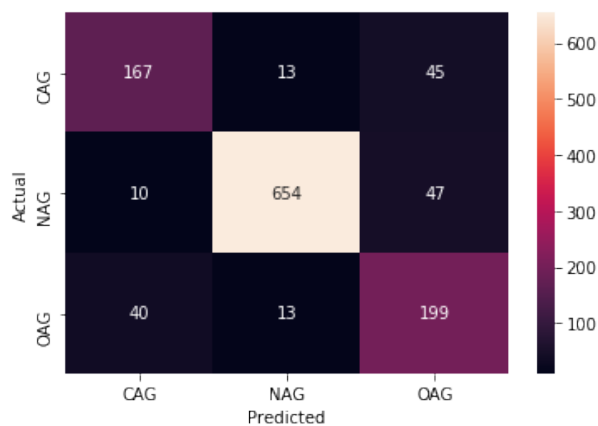

Figure 3: Heat map of the Bengali test set predictions by XLM-R (TL).

Finally, it is clear that in all the experimental settings, the cross-lingual embedding models fine-tuned with transfer learning, outperforms the best system available for the three languages. Furthermore, the results show that the cross-lingual nature of the XLM-R model provided a boost over the multilingual BERT model in all languages tested.

6 Conclusion

This paper is the first study to apply cross-lingual contextual word embeddings in offensive language identification projecting predictions from English to other languages using benchmarked datasets from shared tasks on Bengali (Kumar et al., 2020), Hindi (Mandl et al., 2019), and Spanish (Basile et al., 2019). We have showed that XLM-R with transfer learning outperforms all of the other methods we tested as well as the best results obtained by participants of the three competitions.

The results obtained by our models confirm that OLID's general hierarchical annotation model encompasses multiple types of offensive content such as *aggression*, included in the Bengali dataset, and *hate speech* included in the Hindi and Spanish datasets, allowing us to model different sub-tasks jointly using the methods described in this paper. Furthermore, the results we obtained for Bengali show that it is possible to achieve high performance using transfer learning on off-domain (Twitter vs. Facebook) and off-task data when the labels do not have a direct correspondence in the projected dataset (two in English and three in Bengali). This opens exciting new avenues for future research considering the multitude of phenomena (e.g. *hate speech, aggression, cyberbulling*), annotation schemes and guidelines used in offensive language datasets.

In future work, we would like to further evaluate our models using SOLID, a novel large English dataset with over 9 million tweets (Rosenthal et al., 2020), along with datasets in four other languages (Arabic, Danish, Greek, and Turkish) that were made available for the second edition of OffensEval (Zampieri et al., 2020). These datasets were collected using the same methodology and were annotated according to OLID's guidelines. Finally, we would also like to apply our models to languages with even less resources available to help coping with the problem of offensive language in social media.

Acknowledgement

References

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*.

Rienke Bannink, Suzanne Broeren, Petra M van de Looij-Jansen, Frouwkje G de Waart, and Hein Raat. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. *PloS one*, 9(4).

Md Abul Bashar and Richi Nayak. 2019. QutNocturnal@ HASOC'19: CNN for hate speech and offensive content identification in Hindi language. In *Proceedings of FIRE*.

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*.

Rina A Bonanno and Shelley Hymel. 2013. Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of youth and adolescence*, 42(5):685–697.

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. *arXiv preprint arXiv:1911.02116*.

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. *arXiv preprint arXiv:2005.03695*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of ICLR*.

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. 2020. Evaluating Aggression Identification in Social Media. In *Proceedings of TRAC*.

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*.

Prasenjit Majumder, Thomas Mandl, et al. 2018. Filtering Aggression from the Multilingual Social Media Feed. In *Proceedings TRAC*.

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

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1 – 16.

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, Darwish Kareem, and Magdy Walid. 2017. Abusive language detection on Arabic social media. In *Proceedings of ALW*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

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*.

Juan Manuel Pérez and Franco M Luque. 2019. Atalaya at semeval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of SemEval*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*.

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, 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. *arXiv preprint arXiv:2007.12724*.

Julian Risch and Ralf Krestel. 2020. Bagging bert models for robust aggression identification. In *Proceedings of TRAC*.

Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *arXiv preprint arXiv:2004.14454*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of LREC*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206.

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*.

Luis Enrique Argota Vega, Jorge Carlos Reyes-Magaña, Helena Gómez-Adorno, and Gemma Bel-Enguix. 2019. Mineriaunam at semeval-2019 task 5: Detecting hate speech in twitter using multiple features in a combinatorial framework. In *Proceedings of SemEval*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of ALW*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of NAACL*.

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. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Cagrı Cöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *Proceedings of SemEval*.