UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection

Gregor Wiedemann **Seid Muhie Yimam** **Chris Biemann**
Language Technology Group
Department of Informatics
University of Hamburg, Germany
{gwiedemann, yimam, biemann}@informatik.uni-hamburg.de

Abstract

Fine-tuning of pre-trained transformer networks such as BERT yield state-of-the-art results for text classification tasks. Typically, fine-tuning is performed on task-specific training datasets in a supervised manner. One can also fine-tune in unsupervised manner beforehand by further pre-training the masked language modeling (MLM) task. Hereby, in-domain data for unsupervised MLM resembling the actual classification target dataset allows for domain adaptation of the model. In this paper, we compare current pre-trained transformer networks with and without MLM fine-tuning on their performance for offensive language detection. Our MLM fine-tuned RoBERTa-based classifier officially ranks 1st in the SemEval 2020 Shared Task 12 for the English language. Further experiments with the ALBERT model even surpass this result.

1 Offensive Language Detection

The automatic detection of hate-speech, cyber-bullying, or aggressive and offensive language became a vividly studied task in natural language processing (NLP) in recent years. The offensive language detection (OLD) shared Task 6 of 2019's *International Workshop on Semantic Evaluation* (SemEval) (Zampieri et al., 2019b) attracted submissions from more than 100 teams. The *Offensive Language Identification Dataset* (OLID) used in this shared task comprises three hierarchical classification sub-tasks: A) offensive language detection, B) categorization of offensive language, and C) offensive language target identification (Zampieri et al., 2019a). For Task A, 14,100 Twitter tweets were manually labeled as either *offensive* (OFF) or *not offensive* (NOT). Task B distinguishes the 4,640 offensive tweets from task A into *targeted* (TIN) or general, *untargeted* (UNT) offensive language. Task C, finally, separates targeted insults into the three categories: groups (GRP), individuals (IND), and others (OTH).

OffensEval 2020, the SemEval 2020 *Offensive Language Detection Shared Task*, does not provide an own manually labeled training set for the English language (Zampieri et al., 2020). Instead, a large 'weakly labeled' dataset was published by the organizers, containing roughly nine million tweets. Each tweet has been automatically classified by an ensemble of five different supervised classifiers trained on the OLID dataset. The weakly labeled dataset contains the raw tweet (with user mentions replaced by a special 'USER' token) along with the average label probability and the variance of the five classifier predictions. Since there is no way that such weak labels themselves carry more useful information to a machine learning system than the original dataset on which the five classifiers were trained, we decided not to use any of the weakly labeled information. Instead, for our classification systems, we rely on the 2019 OLID dataset only. However, the OffensEval 2020 dataset is an ample source to build models using unsupervised learning, particularly for domain-adaptation of a pre-trained language model such as BERT (Devlin et al., 2019) or its successors which are based on the transformer neural network architecture. Unfortunately, training a transformer-based language model in an unsupervised manner is incredibly resource-consuming, making it impractical to learn from large datasets without access to larger GPU clusters or TPU hardware. Regarding this, the contribution of our paper is two-fold:

1. We evaluate to what extent different pre-trained transformer-based neural network models can be fine-tuned to detect offensive language and its sub-categories. An ensemble based on the ALBERT (Lan et al., 2019) model achieves the best overall performance.

2. We study how an additional fine-tuning step with masked language modeling (MLM) of the best individual model RoBERTa (Liu et al., 2019b) conducted on in-domain data affects the model performance. An ensemble of models trained with this strategy was submitted as our official contribution to the OffensEval 2020 shared task for the English language and achieved first place in the competition.

2 Related Work

Offensive language detection: Nowadays, a number of public datasets are available to train machine classifiers for detecting English offensive language. Unfortunately, underlying data sources, category definitions, data sampling strategies, and annotation guidelines differ to a large extent between these datasets. Hence, results of different datasets are hardly comparable, and training sets usually cannot be combined to obtain a more robust classification system. Schmidt and Wiegand (2017), and Fortuna and Nunes (2018) conducted insightful surveys on this rapidly growing field. Mandl et al. (2019), Struß et al. (2019), and Basile et al. (2019) recently organized shared tasks on the topic. Although winning systems can achieve striking prediction accuracy, OLD is far from being a solved problem. Prediction performance usually drops severely if the target data comprises different characteristics than the training data. Gröndahl et al. (2018), for instance, show that many machine learning architectures can be fooled easily just by adding the word "love" to an offensive tweet to make it appear as non-offensive. Struß et al. (2019) highlight that linguistic information alone is not enough in many cases to decide whether a tweet is hateful or not. Also context information, e.g. about tweeting users themselves (Ribeiro et al., 2018), or mentioned users in tweets (Wiedemann et al., 2018) can be a useful feature for automatic OLD.

Pre-trained language models for text classification: Transfer learning with deep neural networks, in general, has proven to be superior over supervised learning for text classification, especially for small training data situations. This is illustrated exemplarily in our last year's approach (Wiedemann et al., 2019) to the OLD SemEval shared task which employed unsupervised pre-training of a recurrent neural network architecture with a topic cluster prediction task. Practically all winners of the aforementioned shared task competitions employ some form of a fine-tuned bidirectional transformer-based language model, a neural network architecture for which Devlin et al. (2019) published with BERT the seminal work. This architecture has been proven highly successful for transfer learning. A base model is pre-trained with a MLM task and a next-sentence prediction (NSP) task in an unsupervised manner on very large datasets. The knowledge about language regularities and semantic coherence encoded in the network during this step can then be employed successfully in later training steps of fine-tuning the network weights to the actual classification task. For instance, Liu et al. (2019a) fine-tuned the pre-trained BERT model winning the 2019 SemEval OLD shared task. Also Mozafari et al. (2019), and Risch et al. (2019) used it successfully for offensive language and hate speech detection. Sun et al. (2019) test a wide range of BERT fine-tuning methods for text classification and develop best practice recommendations. Since BERT, a number of successor models improving the network architecture, the pre-training strategy, or the pre-training dataset have been published. A selection of these models will be evaluated in Section 3.

3 Fine-tuning Transformer Networks

We investigate two questions regarding the fine-tuning of pre-trained transformer networks for OLD. First, which pre-trained model performs best on the 2020 OLD shared task? Second, we investigate how much language model fine-tuning on in-domain data prior to classification fine-tuning improves the performance of the best model.

3.1 Model Selection of Transformer Networks

As we have indicated in Section 2, transformer networks have been successfully employed for several text classification tasks. We test the following transformer-based pre-trained models for the OffensEval 2020 OLD shared task.

BERT – Bidirectional Encoder Representations from Transformers: this seminal transformer-based language model employs an attention mechanism that enables to learn contextual relations between (sub-)words in a text sequence (Devlin et al., 2019). BERT uses two training strategies: 1) MLM where 15 % of the tokens in a sequence are replaced (masked) for which the model learns to predict the original tokens, and 2) NSP where the model receives pairs of sentences as input and learns to predict whether or not the second sentence is a successor of the first one in their original document context.

RoBERTa – A Robustly Optimized BERT Pretraining Approach: this is a replication of BERT developed by Facebook (Liu et al., 2019b) with the following modifications 1) training the model longer with bigger batches as well as more and cleaner data, 2) discard the NSP objective, 3) training on longer sequences, and 4) dynamically change the masking patterns, e.g. taking care of masking complete multi-word units. RoBERTa outperformed BERT on most tasks of the GLUE NLP benchmark (ibid.).

XLM-RoBERTa – XLM-R: this is a cross-lingual version of RoBERTa which is trained on several languages at once (Conneau et al., 2019). The model itself is equivalent to RoBERTa, but the training data consists of texts from more than 100 languages filtered from the CommonCrawl[1] dataset.

ALBERT – A Lite BERT for Self-supervised Learning of Language Representations: this is a modification on BERT especially to mitigate memory limitations and training time issues (Lan et al., 2019). The main contributions that ALBERT makes over the design choices of BERT are 1) decomposing the embedding parameters into smaller matrices that will be projected to the hidden space separately, 2) share parameters across layers to improve or stabilize the learned parameters, and 3) inter-sentence coherence loss, which is based on sentence order prediction (SOP), in contrast to BERT's simpler NSP objective.

3.2 Masked Language Model Fine-tuning

Sun et al. (2019) showed that further pre-training of BERT with the masked language model task can improve later results of supervised task-specific fine-tuning. The authors tested *within-task*, *in-domain* and *cross-domain* further pre-training. Evaluations show that the first strategy is susceptible to overfit the training set and, thus, may harm classification performance. The last strategy does not help since BERT is already trained on general-domain data. In-domain further pre-training, however, helps to improve later classification performance if there is a substantial overlap in language characteristics between further pre-training data and supervised training data.

The 'weakly labeled' dataset of the 2020 OLD shared task most certainly is a valuable dataset for further in-domain pre-training. However, with ca. 9 million tweets it is also rather large. Pre-training on the complete dataset is not possible regarding our hardware limitations.[2] Therefore, we conduct MLM pre-training only on a small sample of the original data. We strip URLs and user mentions from tweets, remove duplicates and, finally, randomly sample 5 % of the original dataset size, i.e. 436.123 tweets for further pre-training. We further pre-train the presumably best model *RoBERTa-large* (Liu et al., 2019b) (cp. Section 4) for one epoch (batch size 4, and learning rate 2e-5).

3.3 Ensembling

For our official OffensEval 2020 test set submission as team *UHH-LT*, we aggregated predictions from classifiers with different ensemble approaches.

[1] `https://commoncrawl.org`

[2] MLM of the RoBERTa-large model with the full dataset on a single GPU with 12 GB RAM would take estimated 40 days. However, due to increasing memory consumption of the Adam optimizer during training, the process will stop unfinished way earlier due to a memory exception.

	NOT			**OFF**				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
Baselines								
All NOT	72.21	100.00	41.93	-	0.00	0.00	41.93	72.21
All OFF	-	0.00	0.00	27.78	100.00	43.49	21.74	27.79
Single pre-trained transformer models								
BERT-base	99.06	90.2	94.42	79.34	97.78	87.60	91.01	92.31
BERT-large	99.65	90.35	94.77	79.81	99.17	88.44	91.60	92.80
RoBERTa-base	99.45	90.70	94.88	80.33	98.70	88.57	91.73	92.93
RoBERTa-large	99.53	90.92	95.03	80.73	98.89	88.89	91.96	93.13
XLM-RoBERTa	99.03	91.31	95.01	81.22	97.69	88.69	91.85	93.08
ALBERT-large-v1	98.87	90.24	94.36	79.32	97.31	87.40	90.88	92.20
ALBERT-large-v2	98.87	90.20	94.34	79.26	97.31	87.36	90.85	92.18
ALBERT-xxlarge-v1	98.35	91.09	94.58	80.57	96.02	87.62	91.10	92.46
ALBERT-xxlarge-v2	98.47	91.73	94.98	81.76	96.30	88.44	91.71	93.00
Ensembles of pre-trained transformer models								
All	**99.65**	90.95	95.10	80.83	**99.17**	89.06	92.08	93.23
BERT	99.42	91.16	95.11	81.11	98.61	89.01	92.06	93.23
RoBERTa	99.57	90.84	95.01	80.62	98.98	88.86	91.93	93.11
ALBERT-all	98.23	**92.66**	**95.36**	**83.37**	95.65	89.00	92.23	**93.49**
ALBERT-xxlarge	98.70	92.16	95.32	82.62	96.85	**89.17**	**92.25**	93.47

Table 1: Performance (in %) of baselines, single models, and ensemble models on the OLID test set.

Ensemble of model variants: We fine-tuned different transformer models with the OffensEval 2019 training data using the corresponding test data for validation. The following models were tested: BERT-base and BERT-large (uncased), RoBERTa-base and RoBERTa-large, XLM-RoBERTa, and four different ALBERT models (large-v1, large-v2, xxlarge-v1, and xxlarge-v2). Each model was fine-tuned for 6 epochs with a learning rate of 5e-6, maximum sequence length of 128, and batch size 4. After each epoch, the model was evaluated on the validation set. The best performing epoch was saved for the ensembling. We tested two ensemble approaches: 1) majority vote from all models, and 2) majority vote from one model type but with different parameter sizes such as BERT-base and BERT-large.

MLM RoBERTa ensemble: To be able to learn from the entire 2019 OLID dataset (training and test set), as well as to smooth instabilities of predictions originating from random effects during model training, we also aggregated predictions using 10-fold cross-validation. For this, the further MLM pre-trained RoBERTa-large model is fine-tuned 10 times, each time with 90 % of the OLID data for training and the remaining 10 % as validation set. The best model after 6 epochs of training with learning rate 5e-6 and batch size 8 is used to predict the OLD 2020 test data. The final predictions for submission were obtained via majority vote on the 10 predictions per test data instance.

4 Results

Table 1 shows results of binary offensive language detection for a naive baseline (assuming all tweets as either offensive or not), as well as for the individual fine-tuned transformer models and their corresponding ensembles. All transformer models largely outperform the naïve baseline, some of them (e.g. XLM-RoBERTa) even outperform most of the other system submissions in the competition.[3]

Our best individual model is *RoBERTa-large* with an F1-score of 91.96 %. Hence, we select this model as the basis for further MLM pre-training. From Table 2, we can see that the MLM fine-tuned RoBERTa model achieved consistently better results in terms of Macro F1 than the single pre-trained transformer models (to lower random effects of neural network training, the table shows average values of 10 runs).

[3] `https://sites.google.com/site/offensevalsharedtask/results-and-paper-submission`

	NOT			OFF				
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Macro F1**	**Acc.**
RoBERTa-large	98.96	91.49	95.07	81.54	97.49	88.76	91.93	93.15
RoBERTa-large MLM-ft	**99.15**	**91.53**	**95.18**	**81.66**	**97.96**	**89.06**	**92.12**	**93.31**

Table 2: Performance (in %) of MLM fine-tuned models on the OLID test set (average of 10 runs).

Team	**UHH-LT**	
	Macro F1 (%)	Rank
Task A	92.04	1 out of 84
Task B	65.98	6 out of 42
Task C	66.83	3 out of 38

Table 3: Official results of our OffenseEval 2020 test set submissions for tasks A, B, and C (*RoBERTa-large MLM test* ensemble, cp. Table 2).

Regarding the ensembles of model variants, we see in Table 1 that all approaches consistently perform better than the individual models. Here, the ensemble averaging the predictions from the two *ALBERT-xxlarge* models performed best with an F1-score of 92.25 %.

For the OffensEval 2020 Shared Task, we decided to submit the results form the MLM pre-trained RoBERTa ensemble.[4] Table 3 presents the official results of our system in the sub-tasks A, B, and C together with their ranks achieved in the competition. While our ensemble reaches the top rank for task A, there are a handful of competing approaches achieving better results for offensive language categorization (B) and target identification (C). The post-submission experiments on the official test set as presented in this paper (cp. Table 1) show that the ALBERT-based ensembles would have even beat this first ranked submission. However, the MLM fine-tuned RoBERTa model was considerably more successful on task C than the best ALBERT-xxlarge ensemble, especially to detect the "OTH" class.

Figure 2 shows the corresponding confusion matrices for the submitted predictions. One observation that can be revealed from the matrices is that the predictions for Tasks B and C are considerably biased towards the majority class. For Task A, however, we see more false positive cases for the offensive class which is underrepresented in the training data. A qualitative look into a sample of false predictions (cp. Fig. 1) reveals tweets wrongly predicted as offensive (false positives), some of which seem inconsistently annotated in the gold standard. Parts of expressions in Examples 1, and 2 qualified as offensive in many other training examples. Examples 3, and 4 contain some negative words that may have triggered a higher offensiveness prediction score. For the false negative samples, it is not really obvious why the models missed the correct class, since except for example 6, they actually contain strong offensive vocabulary.

5 Conclusion

After last year's SemEval shared task on offensive language detection was already dominated by the then newly published BERT model, for the year 2020 competition we were successful in fine-tuning BERT's successor models to create the best performing system. The predictions obtained from fine-tuning of the ALBERT model on the OLID dataset achieved 92.25 % macro F1-score as the best overall result (including our post-submission experiments) on the official test set of the Shared Task A for the English language. We also found the 'weak labels' distributed along with 9 million tweets by the shared task organizers not useful for training our classifiers. However, the tweets themselves provided useful in-domain data for unsupervised pre-training. With 92.04 % macro-F1, our predictions based on further language model pre-training of the RoBERTa model on ca. 440.000 tweets, before fine-tuning on last year's OLID dataset achieved the first rank in the official SemEval competition for sub-task A, and also high ranks (6, and 3) for the other two sub-tasks. We conclude that further pre-training of a transformer

[4] Of course, during the submission phase of the shared task, the test set labels were not available. We, thus, based our decision for this specific model on its performance on last year's OLID test set.

1. FP: *@USER the gov and nyc mayor are the biggest joke except maybe for the idiots who elected them*

2. FP: *@USER What the fuck*

3. FP: *@USER I know it was bad but I used to love it*

4. FP: *men who voted yes are childish, sorry are you 17??Men, would you have a problem if a girl said if she's not receiving head she's not giving head?*

5. FN: @USER It's as if every single one of her supporters are just as stupid as her. Wehdone..

6. FN: I'm gonna say he doesn't really. You should check the zip code demographics of his various properties Only liars could deny that Al Sharpton hates white people

7. FN: @USER Fuck the chargers

8. FN: Last night I watched the Democrats throwing shit up against the wall, and none of it stuck.

Figure 1: False positive (FP) and false negative (FN) examples of our UHH-LT Task A submission.

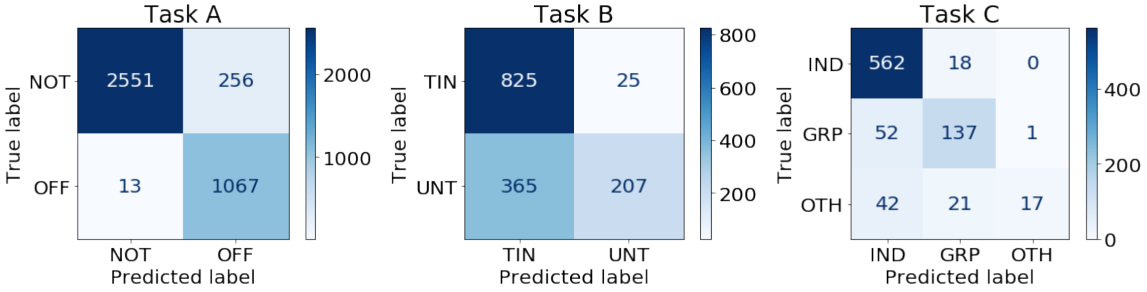

Figure 2: Confusion matrices from the submitted UHH-LT ensemble predictions.

model with in-domain data is useful for offensive language detection. However, for tasks B and C, our models are clearly biased towards the majority class resulting in somewhat lower ranks. Hence, taking the high class imbalance of the OLID dataset better into account could further improve our results.

Acknowledgements

The paper was supported by BWFG Hamburg (Germany) within the "Forum 4.0" project as part of the *ahoi.digital* funding line, and the DFG project "FAME" (WI 4949/2-1, grant no. 406289255).

References

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 the 13th International Workshop on Semantic Evaluation*, pages 54–63, MN, USA. ACL.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, MN, USA. ACL.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, page 2–12, NY, USA. ACM.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, MN, USA, June. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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 the 11th Forum for Information Retrieval Evaluation*, pages 14–17, Kolkata, India.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, *Proceedings of the 8th International Conference on Complex Networks and their Applications*, pages 928–940, Lisbon, Portugal.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *Proceedings of the 12th International Conference on Web and Social Media, ICWSM 2018*, pages 676–679, CA, USA. AAAI Press.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 405–410, Erlangen, Germany.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. ACL.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Germany.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer.

Gregor Wiedemann, Eugen Ruppert, Raghav Jindal, and Chris Biemann. 2018. Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter. In *Proceedings of GermEval Task 2018, 14th Conference on Natural Language Processing (KONVENS)*, pages 85–94, Vienna, Austria.

Gregor Wiedemann, Eugen Ruppert, and Chris Biemann. 2019. UHH-LT at SemEval-2019 task 6: Supervised vs. unsupervised transfer learning for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 782–787, MN, USA. ACL.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420, MN, USA. ACL.

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 The 13th International Workshop on Semantic Evaluation (SemEval)*, MN, USA. ACL.

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 The 14th International Workshop on Semantic Evaluation (SemEval)*, Barcelona, Spain. ACL.