

Galileo at SemEval-2020 Task 12: Multi-lingual Learning for Offensive Language Identification using Pre-trained Language Models

Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, Yu Sun

Baidu Inc., China

{wangshuohuan, liujiaxiang, ouyangxuan, sunyu02}@baidu.com

Abstract

This paper describes Galileo’s performance in SemEval-2020 Task 12 on detecting and categorizing offensive language in social media. For Offensive Language Identification, we proposed a multi-lingual method using Pre-trained Language Models, ERNIE and XLM-R. For offensive language categorization, we proposed a knowledge distillation method trained on soft labels generated by several supervised models. Our team participated in all three sub-tasks. In Sub-task A - Offensive Language Identification, we ranked first in terms of average F1 scores in all languages. We are also the only team which ranked among the top three across all languages. We also took the first place in Sub-task B - Automatic Categorization of Offense Types and Sub-task C - Offence Target Identification.

1 Introduction

Due to the growing number of Internet users, cyber-violence emerged with offensive language pervasive across social media. With anonymity as a “privilege”, netizens hide behind the screens, behaving in a manner most of them would not otherwise in reality. Thus, government organizations, online communities, and technology companies are all striving for ways to detect aggressive language in social media and help build a more friendly online environment.

Manual filtering is very time consuming and it can cause post-traumatic stress disorder-like symptoms to human annotators. One of the most common strategies (Waseem et al., 2017; Davidson et al., 2017; Malmasi and Zampieri, 2018; Kumar et al., 2018) to tackle the problem is to train systems capable of recognizing offensive content, which can then be deleted or set aside human moderation.

SemEval 2020 Task-12 (Zampieri et al., 2020) is the second edition of OffensEval (Zampieri et al., 2019). In this competition, organizers offers 5 languages datasets including Arabic (Mubarak et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), English (Rosenthal et al., 2020), Turkish (Çöltekin, 2020) and Greek (Pitenis et al., 2020). In Sub-task A, the participants need to predict whether a post uses offensives language. Besides, the organizers provide other two sub-tasks which mainly focus on English, to predict the type and target of offensive language.

Participating in all 3 Sub-tasks, we proposed several methods based on pre-training language models including ERNIE and XLM-R. In Sub-task A, we scored 0.9199, 0.851, 0.8258, 0.802, 0.8989 in English, Greek, Turkish, Danish and Arabic respectively. We ranked first in average F1 scores, and ranked in top three across all languages. In Sub-task B and Sub-task C, we also took the first place with 0.7462 and 0.7145. In the following sections, we will elaborate the methods, dataset and experiments of our system.

2 Related Work

2.1 Monolingual Pre-trained Language Models

Pre-training and fine-tuning have become a new paradigm in natural language processing, where the general knowledge is firstly learnt from large-scale corpus through self-supervised learning and then

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

transferred to down-stream tasks for task-specific fine-tuning. The following are some representatives.

(Peters et al., 2018) proposed context-sensitive word vectors (ELMo) that enhance downstream tasks by acting as features. (Radford et al., 2018) proposed GPT which enhanced the context-sensitive embedding by adjusting the Transformer (Vaswani et al., 2017). (Devlin et al., 2019) modeled a bidirectional language model (BERT) through a task similar to Cloze. (Yang et al., 2019) proposed a permuted language model (XLNet) which is a generalized autoregressive pre-training method. (Liu et al., 2019b) remove the next prediction task and pre-train longer to get a better pre-trained model (RoBERTa). (Clark et al., 2019) proposed a method to joint generator and discriminator in ELECTRA. (Lan et al., 2019) and (Raffel et al., 2019) explored the larger model structure while optimizing the pre-training strategy in ALBERT and T5.

(Sun et al., 2019) enhanced pre-trained language models with full masking of spans in ERNIE. (Sun et al., 2020) proposed continuous multi-task pre-training and several pre-training tasks in ERNIE 2.0. The researchers of ERNIE 2.0 released a new version recently which made a few improvements on knowledge masking and application-oriented tasks, with the aim to advance the model’s general semantic representation capability. In order to improve the knowledge masking strategy, they proposed a new mutual information based dynamic knowledge masking algorithm. They also constructed pre-training tasks that are specific for different applications. For example, they added a coreference resolution task to identify all expressions in a text that refer to the same entity. For more details, please go to this website¹.

2.2 Cross-lingual Pre-trained Language Models

In addition, there are also a lot of works on multilingual language models. (Devlin et al., 2019) provided a multilingual version of BERT that demonstrates surprising cross-language capabilities (Wu and Dredze, 2019). (Conneau and Lample, 2019) proposed two tasks, Masked Language Model and Translation Language Model, to model monolingual corpus and bilingual parallel corpus respectively. (Huang et al., 2019) proposed Unicoder incorporate more bilingual parallel corpus modeling methods. (Song et al., 2019) and (Liu et al., 2020) proposed modeling methods that are more suitable for machine translation tasks in MASS and MBART. (Conneau et al., 2019) used the ideas of RoBERTa in XLM-R and achieved better results than XLM.

2.3 Methods of Offensive Language Detection and Categorization

In the last few years, there have been several studies on the application of computational methods to cope with offensive language. (Waseem et al., 2017) proposed a typology that captures central similarities and differences between subtasks. (Davidson et al., 2017) trained a multi-class classifier to distinguish between these different categories. (Malmasi and Zampieri, 2018) employed supervised classification along with a set of features that includes n-grams, skip-grams and clustering-based word representations. There are also several workshops for this problem. Such as AWL² and TRAC³ (Kumar et al., 2018).

Besides, there are several works focus on offensive language identification in languages other than English, such as Chinese (Su et al., 2017), Dutch (Tulkens et al., 2016), German (Ross et al., 2016), Slovene (Fišer et al., 2017) and Arabic (Mubarak et al., 2017). There are also some researches (Basile et al., 2019; Mandl et al., 2019) about multilingual offensive language identification.

3 Methodology

3.1 Multi-lingual Offensive Language Detection

In Sub-task A, we expected to build a unified approach to detect offensive language in all languages.

Our algorithm has two steps. In the first step, pre-training using large scale multilingual unsupervised texts yields a unified pre-training model that can learn all the language representations together. In the second step, the pre-trained model was fine-tuned with labeled data. The detailed process is shown in Figure 1.

¹<http://research.baidu.com/Blog/index-view?id=128>

²<https://sites.google.com/view/alw3/>

³<https://sites.google.com/view/trac1/home>

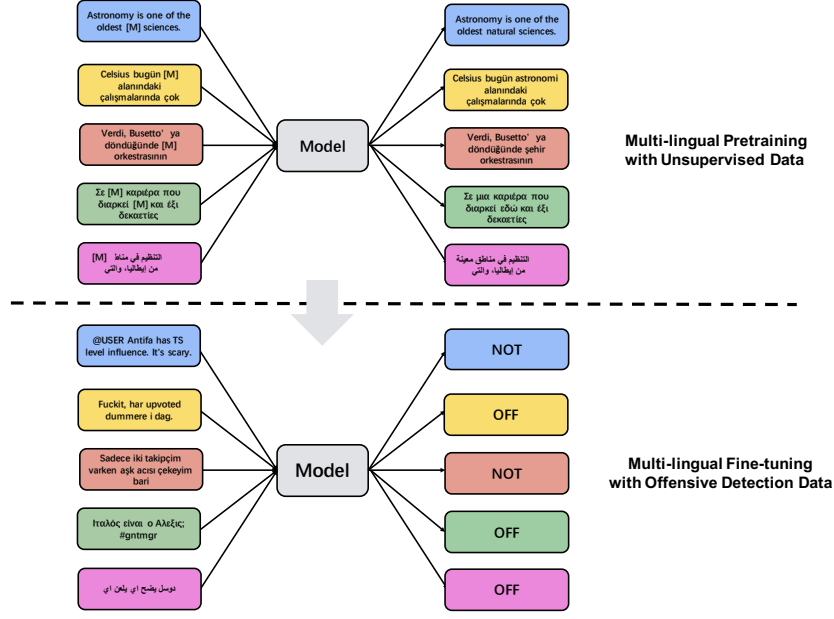


Figure 1: The Framework of Multi-lingual Offensive Language Detection

To skip large-scale pre-training, we used an existing open-source model, XLM-R, as our first step. With Transformer as the backbone structure, XLM-R was pretrained with masked language model on Common Crawl dataset in over 100 languages.

We add a full connected layer for classification upon the [CLS] position of the top layer of XLM-R, using the same parameter for all languages.

This approach can benefit from dataset in other languages and enhance the generality of the model. We will compare methods trained on multilingual data with those on monolingual data in Section 5.

3.2 Offensive Language Categorization using Knowledge Distillation trained on Soft Labels

In Sub-task B and Sub-task C, we constructed a knowledge distillation approach (Hinton et al., 2015; Davidson et al., 2017; Liu et al., 2019a). Several supervised models provided calculated the probability of each label and generated a weighted probability (here we call it soft label). Then the student model was trained on those soft labels. Detailed process is shown in Figure 2.

Suppose that X is the contextual embedding of the token [CLS], which can be viewed as the semantic representation of input sentence. Let $Q(c|X)$ be the class probabilities produced by the ensemble of several supervised models. The probability $P_r(c|X)$ that X is labeled as class c is predicted by a softmax layer. We use the standard cross entropy loss to learn the soft target:

$$Loss = - \sum_c Q(c|X) \log(P_r(c|X)) \quad (1)$$

We used ERNIE 2.0 and ALBERT as our candidates of pre-training language models in Sub-task B and Sub-task C.

4 Dataset

We used datasets of OffensEval 2019 and OffensEval 2020 as our training data. In OffensEval 2019, the organizers provide a dataset containing English tweets annotated using a hierarchical three-level annotation. In OffensEval 2020, the organizers did not provide additional data in English for training. They provided training data for four other languages, Turkish, Danish, Greek and Arabic. In addition, they provide a large amount of weakly labeled data generated by several supervised models.

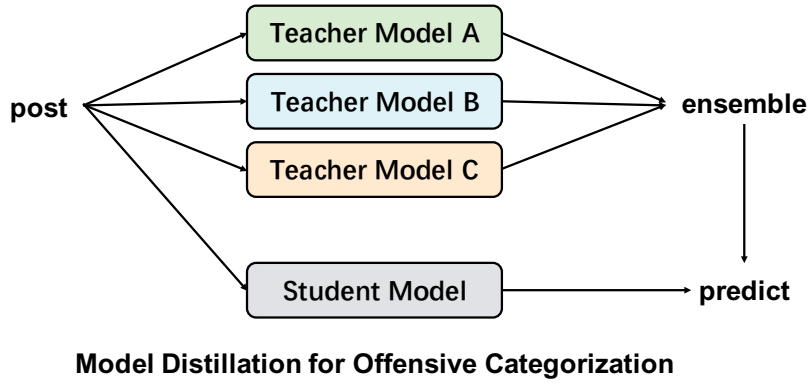


Figure 2: The Framework of Offensive Language Categorization

4.1 Sub-Task A - Offensive Language Identification

In Sub-task A, the goal is to discriminate between offensive and non-offensive posts. Offensive posts include insults, threats, and posts containing any form of untargeted profanity. Each instance is assigned one of the following two labels. 'NOT' means posts which do not contain offense or profanity. 'OFF' means posts containing offense any form of non-acceptable language or a targeted offense.

In order to avoid uneven proportions of data across languages, we did not use the unannotated English data from OffensEval 2020. Instead, we used a mix of English data from OffensEval 2019 (including training data and test data) and training data from OffensEval 2020 in the other 4 languages as our training data. Details are shown in Table 1.

Languages	Train			Test		
	OFF	NOT	TOTAL	OFF	NOT	TOTAL
English	4640	9460	14100	1080	2807	3887
Turkish	6131	25625	31756	716	2812	3528
Arabic	1589	6411	8000	402	1598	2000
Danish	384	2577	2961	41	288	329
Greek	2486	6257	8743	242	1302	1546

Table 1: Dataset Statistics for Sub-task A

4.2 Sub-Task B - Automatic Offense Language Categorization

In Sub-task B, the goal is to predict the type of offense. There are two types in sub-task B are the following. 'TIN' means posts containing an insult or threat to an individual, group, or others. 'UNT' means posts containing non-targeted profanity and swearing. The dataset consists of two parts, a small portion of the manually annotated dataset from OffensEval 2019 and a large portion of the dataset from OffensEval 2020 constructed based on multiple supervision models. All the training data in OffensEval 2020 provides the confidence that it has a target to attack. Details are shown in Table 2.

DataSet	Train			Test		
	TIN	UNT	TOTAL	TIN	UNT	TOTAL
OffensEval 2019	4089	551	4640	-	-	-
OffensEval 2020	149550	39424	188974	850	572	1422

Table 2: Dataset Statistics for Sub-Task B

4.3 Sub-Task C - Offense Target Identification

In Sub-Task C, the goal is to predict the target of offense. The three labels in Sub-task C are the following. 'IND' means posts targeting an individual. 'GRP' means the target of these offensive posts is a group of people. 'OTH' means the target of these offensive posts does not belong to any of the previous two categories. As with Sub-task B, all training data in OffensEval 2020 provide the confidence level for each label. Details are shown in Table 3.

DataSet	Train				Test			
	IND	GRP	OTH	TOTAL	IND	GRP	OTH	TOTAL
OffensEval 2019	2507	1152	430	4089	-	-	-	-
OffensEval 2020	152562	24917	11494	188973	580	190	80	850

Table 3: Dataset Statistics for Sub-Task C

5 Experiments

5.1 Results of Sub-task A

We validated our proposed methods based on two models, XLM-R Base and XLM-R Large. The metric used is the average F1 score of all labels. To make the results more reliable, we repeated the experiment 5 times and used the average F1 score. In Table 4, we can see that the result of multilingual fine-tuning is better in all languages except Turkish. It might be caused by it taking up the largest proportion among all languages, leading to data of other languages being ignored. In the table below, we also listed the final submitted results and ranks in the contest, where we used ten-fold cross-validation-based ensemble of XLM-R Large.

Languages	XLM-R BASE		XLM-R LARGE		Submitted Result(Ensemble)	
	Single	Multi	Single	Multi	Multi	rank in all teams
English	0.9150	0.9214	0.9186	0.9255	0.9199	3
Turkish	0.8081	0.8084	0.8265	0.8224	0.8258	1
Arabic	0.8649	0.8730	0.8969	0.9015	0.8989	3
Danish	0.7733	0.7922	0.7908	0.8136	0.8020	2
Greek	0.8266	0.8356	0.8356	0.8392	0.8510	2
Average	0.8376	0.8461	0.8537	0.8604	0.8595	1

Table 4: Results for Sub-Task A. We report the F1 score of all languages.

5.2 Results of Sub-task B and Sub-task C

In both Sub-Task B and Sub-Task C, we made a comparison between hard target-based approach and soft target-based approach. Two models were used for validation, which are ALBERT-XXLarge and ERNIE 2.0. The results are shown in Table 5 and Table 6, where it can be seen that the knowledge distillation approach is helpful for offensive categorization.

Same with Sub-Task A, the metric used is the average F1 score of all labels. Again, to make it more reliable, the average score of 5 repeated experiments was adopted. We also listed our final submitted results below, which were obtained using ten-fold cross-validation-based ensemble of ERNIE 2.0.

6 Conclusion

In this paper, we presented our approach on detecting and categorizing offensive language in social media. We proposed a multi-lingual learning method to detect offensive language and a knowledge

Method	ALBERT(F1)	ERNIE 2.0(F1)
Learning with Hard Target	0.6810	0.6883
Learning with Soft Target	0.7043	0.7124
Submitted Result	-	0.7462

Table 5: Results for Sub-task B

Method	ALBERT(F1)	ERNIE 2.0(F1)
Learning with Hard Target	0.6727	0.6773
Learning with Soft Target	0.6864	0.6894
Submitted Result	-	0.7145

Table 6: Results for Sub-task C

distillation method to categorize offensive language. We will further our exploration of multilingual offensive language identification in future, e.g. validating the zero-shot performance of our model in more languages.

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.
- Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. ELRA.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.
- 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*.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- 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 of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Uni-coder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.
- Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

- 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. In *International Conference on Learning Representations*.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.
- 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. *arXiv preprint arXiv:1907.11692*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.
- Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.
- 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.
- Hamdy Mubarak, Darwish Kareem, and Magdy Walid. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.
- 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*.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.
- Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In *arxiv*.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC)*, Bochum, Germany.
- Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*.
- Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.
- Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975.
- Stéphan Tulkens, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A Dictionary-based Approach to Racism Detection in Dutch Social Media. In *Proceedings of the Workshop Text Analytics for Cybersecurity and Online Safety (TA-COS)*, Portoroz, Slovenia.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Zeeraq Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.
- Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.
- 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*.