ANDES at SemEval-2020 Task 12: A jointly-trained BERT multilingual model for offensive language detection

Juan Manuel Pérez[*]
ICC, CONICET
Universidad de Buenos Aires
jmperez@dc.uba.ar

Aymé Arango[*]
University of Chile
IMFD, Chile
aarango@dcc.uchile.cl

Franco M. Luque
CONICET
Universidad Nacional de Córdoba
francolq@famaf.unc.edu.ar

Abstract

This paper describes our participation in SemEval-2020 Task 12: Multilingual Offensive Language Detection. We jointly-trained a single model by fine-tuning Multilingual BERT to tackle the task across all the proposed languages: English, Danish, Turkish, Greek and Arabic. Our single model had competitive results, with a performance close to top-performing systems in spite of sharing the same parameters across all languages. Zero-shot and few-shot experiments were also conducted to analyze the transference performance among these languages. We make our code public for further research[1].

1 Introduction

Offensive language, hate speech, cyberbullying, and abusive language detection are topics that have summoned a lot of interest in the last years, particularly due to the necessity in Social Media to stop –or at least, to diminish– this omnipresent phenomenon. Not only the social implications are behind this, but also practical implications for companies: recently, some large advertisers have removed their presence from Social Media as they consider that platforms full of "divisiveness and hate speech" do not give value to their companies (Hern, 2020).

The differences between these categories are loosely defined. Waseem et al. (2017) propose using two axes to understand the particularities of each discourse: whether the abusive language is directed to a specific individual or to a generalized group, and whether the offending discourse is explicit or implicit.

Even when online platforms prohibit behavior that crosses the line into abuse,[2] these rules are frequently violated. Automated moderation algorithms are necessary to perform a faster and even better user-generated-content moderation or to serve as a tool that helps human moderators to reduce the volume of offensive content still present in online platforms. To address this problem, offensive language detection is usually thought of as a binary classification problem in which the input is a post (tweet, message, comment, etc.) and the output is the classification of whether the post is offensive or not.

Several features of offensive language make it complex to detect. The *offenders* could intentionally hide offensive words by substituting letters with special characters or numbers. Content with irony or sarcasm could be harmful even when all words are polite, and vice versa. The real intention behind speech is sometimes difficult to detect even for humans.

According to Chatzakou et al. (2017), in the case of Twitter, each tweet provides a fairly limited context. Therefore, an offensive post may be identified as unoffensive if the context is not taken into account. Despite several published proposals on detecting offensive language, hate speech, and other related concepts, there is no consolidated or effective strategy that could be applied to different languages and domains.

---

[*]Authors contributed equally
[1]https://github.com/finiteautomata/offenseval2020

[2]https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content

Building a quality labeled dataset is an expensive task in time and effort. Most available datasets are in English, and often datasets are not published due to privacy concerns, making research on offensive language detection difficult for other languages. The generalization power of learning models would allow us to transfer knowledge to languages with poor resources.

In this work, we describe our participation in SemEval 2020 Task 12: OffensEval, an offensive language detection task in five different languages. We propose a single multilingual system based on Multilingual BERT (Devlin et al., 2018) to jointly address the offensive language detection. We trained our system using all the available training data and evaluated its performance for each of the languages instead of adjusting language-specific models. Our single model performs fairly well, achieving performances comparable to the winning teams for each language. To further explore the multilingual dimension of BERT, we analyze zero-shot and few-shot capabilities in a cross-lingual setting for the task in question. We make our code public for further research.

2 Background

The detection of offensive language, cyberbullying and hate speech are tasks that are closely connected and often confused (Malmasi and Zampieri, 2018). Several machine learning models addressing hate speech or offensive language detection have been proposed in the last years; in particular *deep learning* models (Gambäck and Sikdar, 2017; Park and Fung, 2017; Badjatiya et al., 2017; Agrawal and Awekar, 2018; Bisht et al., 2020; Gertner et al., 2019; Pérez and Luque, 2019) have increased their popularity among researchers on this task. Despite the growing interest in the area, the models are usually trained and evaluated inside very specific English datasets, and their generalizability to other contexts or languages is still a challenge. Morever, building these datasets is difficult (Waseem, 2016) and achieved performances are often overestimated (Arango et al., 2019; Wiegand et al., 2019).

There are only a few studies addressing multilingual detection of these subtypes of abusive language in the related literature. In these, authors proposed single systems that can be used to classify data in different languages. Some of the common features used in this kind of models are multilingual word representations such as MUSE (Conneau et al., 2017) or Multilingual BERT (Devlin et al., 2019). Other authors combined word embeddings with tweet-level features (Corazza et al., 2020) or linguistic features (Benito et al., 2019).

In most cases, the multilingual models are trained and tested independently for each language and do not combine different languages in a single evaluation. An exception is the approach proposed by Bojkovský and Pikuliak (2019), where the authors trained deep neural network architectures with a concatenation of English and Spanish datasets to classify data in both languages.

Although Multilingual BERT models have been tested as end-to-end solutions for several tasks, they have not been widely explored for offensive language detection. Pires et al. (2019) tested the zero-shot capability of BERT for transferring knowledge from one language to another in *named entity recognition* and *part of speech tagging* tasks obtaining high performing results.

3 Data

The dataset used in this work is described thoroughly in Zampieri et al. (2020). We decided to replace the distantly-supervised English training dataset(Rosenthal et al., 2020) presented in this task by the OLID dataset of Zampieri et al. (2019). This is because the focus of this work is the multilingual capabilities of BERT, and the datasets of the languages added to this task (Danish(Sigurbergsson and Derczynski, 2020), Greek(Pitenis et al., 2020), Arabic(Mubarak et al., 2020) and Turkish(Çöltekin, 2020)) resemble more the OLID dataset in that they were manually annotated.

4 System Overview

Our model is a fine-tuned version of Multilingual BERT (Devlin et al., 2018). This architecture (which has become the state of the art for most NLP tasks) consists of a stack of transformer blocks (Vaswani et al., 2017) pretrained in two tasks: masked language model (also known as Cloze task) and next sentence

Language	Train	Dev	Test
English	13240	860	3887
Greek	6994	1749	1544
Danish	2368	592	329
Arabic	6839	1000	2000
Turkish	25021	6256	3528

Table 1: Size of the train, development and test datasets used in this work.

prediction. For the downstream task an output layer is added and the model is fine-tuned with very low learning rates.

Multilingual BERT shares the same training as single-language BERT but using a concatenated dataset of 104 languages, and has demonstrated to have surprising cross-lingual capabilities, even among languages that do not share scripts (Pires et al., 2019).

The implementation used in this work is the pretrained multilingual BERT-base model from the HuggingFace library(Wolf et al., 2019). This model consists of 12 transformer blocks, 12 self-attention heads, and a hidden layer size of 768. On the top of it, we added a linear layer and applied a sigmoid function to the outputs.

We trained the model for 10 epochs with a batch size of 32 using a dropout probability of 0.1, setting the initial learning rate at $5 * 10^{-5}$ and binary cross entropy as the loss function. Adam with linear warm-up of 10% of the steps was used to optimize the loss.

5 Experimental Setup

We tested several experimental configurations using the data described in Table 1. The main purpose of our different setups is to test the capability of multilingual models not only in inside-language evaluation but also generalizing knowledge from one language to another. The generalization capability of offensive language detection models across different languages has been poorly explored.

For monolingual evaluation, we trained our model using each one of the training sets and the corresponding validation and testing sets. This experimental setup allows us not only to test the model in a specific language but also to obtain reference values to be compared to the ones obtained in the multilingual experimental setups. We refer to this setting as *BERT Lang* (*BERT Greek*, *BERT English*, etc).

As a second configuration, we opted for multilingual training. We trained our model with the concatenation of all the training sets and evaluated it over each test set. The purpose of this experiment is to find languages that contribute positively to the monolingual classification. We call this setting *BERT All*.

It might be argued that in monolingual settings it would have been a better option to simply use the *BERT* version trained specifically for that language. However, we decided to use the multilingual version to have comparable results.

To assess the multilingual potential of BERT for this task, we also performed some *zero-shot* and *few-shot* experiments. Zero-shot experiments consisted in training the model in one language and evaluating it in a different one. That is, training with language *A* and testing with language *B*. Few shot experiments, on the other hand, trains the model in language *A* using also a little amount of instances from *B*, and tests their performance against *B*. This cross-lingual generalization is desirable to tackle the same problem in low-resourced languages. In Section 5 we discuss the results in each case.

The evaluation metric proposed for this task is *Macro F1*.

5.1 Error Analysis and Model Interpretation

To analyze the reasons behind the errors of our model, we used the Captum library (Kokhlikyan et al., 2019) implementation of Integrated Gradients (Sundararajan et al., 2017) to have more information about the importance of each token towards classifying the tweet as offensive. This returns, for a model and a

Model	English	Danish	Greek	Arabic	Turkish
Baseline	0.720	0.690	0.639	0.547	0.452
BERT English	**0.895**	0.646	0.573	0.473	0.555
BERT Danish	0.713	**0.740**	0.473	0.473	0.463
BERT Greek	0.436	0.499	**0.840**	0.446	0.498
BERT Arabic	0.432	0.486	0.502	**0.859**	0.490
BERT Turkish	0.431	0.501	0.499	0.545	**0.766**
BERT All	**0.899**	**0.772**	**0.815**	**0.840**	**0.773**
Best System	0.922	0.812	0.852	0.901	0.825
Approx diff	0.023	0.040	0.037	0.061	0.052

Table 2: Performance of each model for the different datasets. BERT Lang refers to BERT using that language as training data. The Baseline model is a bidirectional LSTM trained and tested in the same language. Performance is measured in Macro-F1 score. Non-diagonal entries show zero-shot performances. Bold letters indicate top two results.

Model	Danish
BERT (Danish + English)	0.733
BERT (Danish + Greek)	0.720
BERT (Danish + Turkish)	0.697
BERT (Danish + Arabic)	**0.792**

Table 3: Performance on data augmentation experiments over the Danish test set. BERT Lang refers to BERT using that lang as training data. Performance is measured in Macro-F1 score.

sentence, a value for each token representing the contribution of it towards the positive class (offensive) or towards the negative class (not offensive).

6   Results

Table 2 shows the results for each of our trained classifiers. The classifier presented for the competence is *BERT All*, in spite of having better performing systems in the monolingual settings; for instance, *BERT Greek* has better results for Greek than *BERT All*. However, *BERT All* performs fairly well, having small differences with the best performing system for each language. For most languages, it stays in the "top cluster" of the competition for each language, most notably in Danish achieving the 7th position. We must remember that *BERT All* is a single model for all the languages, reducing the need for several models – in our case, a reduction of five-to-one.

In most cases, *BERT All* showed results equal or slightly worse than the monolingual setting, telling us (at first sight) that adding other languages does not contribute to the overall performance. In the case the Turkish language, however, there is a slight increase from 0.766 to 0.773 of F-score. More interesting is the case of the Danish dataset, where the F-score increased from 0.74 to 0.77.

To find out the impact of the other languages over the Danish results, we conducted a data augmentation experiment. We augmented the training Danish dataset with other languages data to classify the tweets in the Danish dataset. Table 3 shows the results of this experiment. The results are, in general reasonable. Adding data from different languages does not dramatically impact the monolingual results. The addition of the Arabic dataset turned out to be the most successful one, despite having been the worst result in the zero-shot experiment. It is somehow surprising that a language from a very different family might positively impact in the performance of the Danish classifier.

Regarding the zero-shot cross-lingual capability (the non-diagonal entries in Table 2) it can be observed that there is no transfer learning in this mode. That is, no classifier trained in one language performs successfully when tested in a different one. The only exception for this is Danish-English, as we can see
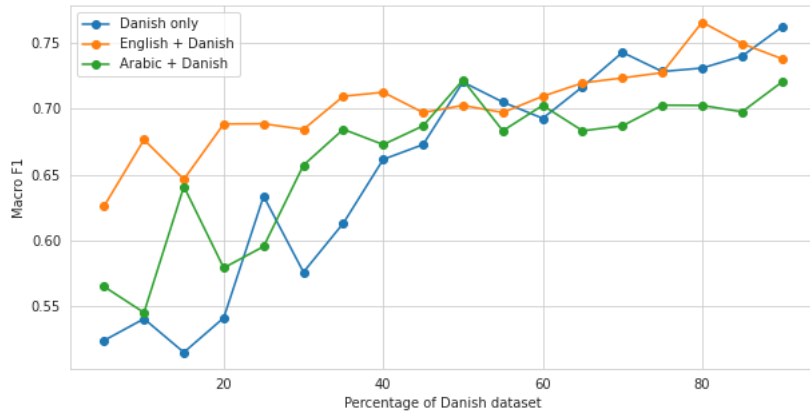
Figure 1: Results of Few-shot training for Danish. Blue line shows the performance of the classifier trained with the respective percentage of the training dataset, and the orange line corresponds to the classifier trained with English plus the respective percentage of the Danish dataset.

(a) Confusion Matrix of BERT All in English

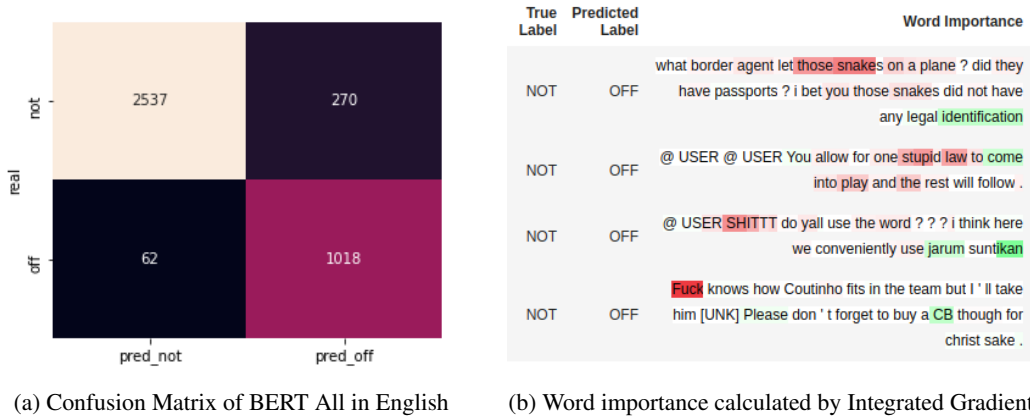(b) Word importance calculated by Integrated Gradients

Figure 2: Error Analysis for our model. Figure a displays the confusion matrix of BERT All in English. Figure b displays the word importance for each token: red means that the token is pushing towards the "offensive" class, whereas green pulls to not offensive class.

that training with English and testing against Danish gives something around 0.65 Macro F1. However, a close examination of the true positives in this case brings us to the conclusion that the transference is mainly due to vocabulary sharing and code switching.

Few-shot experiments yielded slightly better results. Figure 1 displays the performance of classifiers trained with $5\%, 10\%, 15\%\dots$ of the instances of the Danish dataset (blue line) and the performance of training the same classifier but using also another dataset: English or Arabic. Performance using 20% of the Danish data and the English dataset achieves 0.69 Macro-F1 score while we need almost the double of data to have the same performance using Danish-only training. Using Arabic dataset yields marginal improvement in performance.

6.1 Error Analysis

Figure 2a displays the confusion matrix for our model against the English test set. We can observe that most errors come from false positives. Figure 2b shows the word importance for a couple of false positive examples. It turns out that numerous spurious correlations are learned by our classifier: words such as "Trump", "disgusting", "racist" trigger the sentence as offensive. Also, more complex constructions such as reporting offensive incidents are not understood by the classifier: for instance, "They call me b*tch" should not be marked as offensive.

It is also important to notice that there are a number of instances that seem to be mislabelled. For instance, these examples were labelled as non-offensive:

- what border agent let those snakes on a plane? did they have passports? i bet you those snakes did not have any legal identification

- It's Me and IDGAF About Nothing – Females with nigga mindsets are dangerous

- Fuck knows how Coutinho fits in the team but I'll take him Please don't forget to buy a CB though for christ sake.

Whereas these are considered offensive:

- Crazy that as we get older and go through certain shit you just want to keep it to yourself.

- @USER Trump lied..and then coldly said it now appeared his building was the tallest in NYC. He's a sick, sick twist.

- Knew sis was a liar but I got soft anyways smh

- It's crazy how people make excuses for them to walk out of a person's life...

In some cases it seems that there are instances that are wrongly labelled, and in other cases there are inconsistencies regarding swearing – is saying "sh*t" or "f*ck" considered offensive? It's known that offensive language (alongside with cyberbullying, hate speech and so on) are difficult tasks for the annotators and that non-skilled annotators deliver datasets having a lot of noise.

7 Conclusion

In this work, we explored the capabilities of Multilingual BERT for offensive language detection in a multilingual scenario. We evaluated our model in different experimental setups: training and testing it individually for each language; training with one language data and testing in others (zero-shot mode); and training a single model on all languages.

Something to notice is that, in spite of the good performance of multilingual BERT in zero-shot cross-lingual evaluation for other tasks, it did not work well for this one. The only exception is the English-Danish cross-lingual evaluation, and it is mainly due to vocabulary intersection. Further work is needed to analyze the reasons why these zero-shot experiments failed. Nonetheless, the only experiment performed in few-shot mode showed slightly better results, letting a Danish model with just a hundred of examples achieve a reasonable performance.

On the other hand, training all the languages jointly resulted in a single model having a similar performance overall. This is interesting as having a single model instead of five has practical implications, in particular concerning the resources needed for big models such as BERT.

The relationship of offensive language among different languages is something to be studied in more depth. At first glance, our experiments showed no zero-shot transference; a few-shot experiment showed (in the case of two languages of similar typology) some transference. Further study is needed to explain these results.

Acknowledgements

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 141–153.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 45–54.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Diego Benito, Oscar Araque, and Carlos Angel Iglesias. 2019. GSI-UPM at semeval-2019 task 5: Semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 396–403.

Akanksha Bisht, Annapurna Singh, HS Bhadauria, Jitendra Virmani, et al. 2020. Detection of hate speech and offensive language in Twitter data using lstm model. In *Recent Trends in Image and Signal Processing in Computer Vision*, pages 243–264. Springer.

Michal Bojkovský and Matús Pikuliak. 2019. STUFIIT at Semeval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and Elmo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 464–468.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 6174–6184. ELRA.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 13–22.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics.

Abigail S Gertner, John Henderson, Elizabeth Merkhofer, Amy Marsh, Ben Wellner, and Guido Zarrella. 2019. Mitre at semeval-2019 task 5: Transfer learning for multilingual hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 453–459.

Alex Hern. 2020. How hate speech campaigners found facebook's weak spot.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Jonathan Reynolds, Alexander Melnikov, Natalia Lunova, and Orion Reblitz-Richardson. 2019. Pytorch captum. https://github.com/pytorch/captum.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

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*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics.

Juan Manuel Pérez and Franco M Luque. 2019. Atalaya at semeval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In *arxiv*.

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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

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.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

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*.