GruPaTo at SemEval-2020 Task 12: Retraining mBERT on Social Media and Fine-tune Offensive Language Models

Davide Colla[♣◇], **Tommaso Caselli**[♣], **Valerio Basile**[◇], **Jelena Mitrović**[‡], **Michael Granitzer**[‡]
[♣]University of Groningen, [◇]University of Turin, [‡]University of Passau
Groningen The Netherlands, Turin Italy, Passau Germany
◇{valerio.basile|davide.colla}@unito.it, ♣t.caselli@rug.nl
‡jelena.mitrovic|michael.granitzer}@uni-passau.de

Abstract

We introduce an approach to multilingual Offensive Language Detection based on the mBERT transformer model. We download extra training data from Twitter in English, Danish, and Turkish, and use it to re-train the model. We then fine-tuned the model on the provided training data and, in some configurations, implement transfer learning approach exploiting the typological relatedness of the English and Danish languages. Our systems obtained good results across the three languages (.9036 for EN, .7619 for DA, and .7789 for TR).

1 Introduction

The growth of Social Media has seen the spread of two different but connected phenomena: on the one hand, they helped to create a more open and connected world, and, on the other hand, they contributed to the spread of offensive and abusive behaviors. Although the use of "bad language" is intimately connected with freedom of speech, the phenomenon has become so pervasive that developing Natural Language Processing (NLP) systems that automatically and efficiently detect and classify offensive on-line content is a pressing need (Nobata et al., 2016; Kennedy et al., 2017).

SemEval-2020 Task 12: OffensEval 2 (Zampieri et al., 2020) is a follow-up edition of SemEval-2019 Task 6: OffensEval (Zampieri et al., 2019a) and it addresses the problem of offensive language detection in Twitter messages by focusing on two pending issues: multilingualism and hierarchical tagset annotation.

The multilingualism issue is targeted by providing for the first time data in 5 different languages, namely English, Danish, Greek, Turkish, and Arabic, by applying a shared definition of offensive language. The languages cover different values of the typological spectrum in terms of type (Fusional *vs.* Agglutinative), language family (Indo-European *vs.* Altaic *vs.* Afro-African), genus (Germanic *vs.* Greek *vs.* Turkic *vs.* Semitic), Subject-Object-Verb [1] order (SVO *vs.* no dominant order *vs.* SOV *vs.* VSO) (Dryer and Haspelmath, 2013; Ramat and Baldry, 2011), as well as writing systems. The multilingual aspect poses two additional challenges: (i.) availability of NLP tools and language resources as some of the proposed languages are considered low-resourced (e.g., Danish and Greek) (Rehm and Uszkoreit, 2013); and (ii.) differences in the perceived offensiveness of a message. In particular, given that offensiveness is a highly subjective category, seeing that a message is always "offensive for someone" (Vidgen et al., 2019), different communities of speakers may have different perception of what is offensive or not. The use of a shared definition is a way of mitigating the potential differences across communities, but this aspect cannot be ignored in the development of a system for offensive language detection.

The hierarchical annotation tagset is reflected in three sub-tasks, namely:

- Sub-task A: Offensive language identification: The task consists in predicting if a tweet is offensive or not. The definition of "offensive message" (OFF) is based on the SemEval-2019 Task 6 (Zampieri

[1]Generally, subject and object are used in an informal semantic sense to denote the more agent-like and more patient-like elements (Dryer, 2013)

et al., 2019b), namely "[p]osts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words." (Zampieri et al., 2019a, p. 1416)

- Sub-task B: Automatic categorization of offense types: The task consists in predicting the type of offense. It applies only to messages labelled as offensive (OFF) in sub-task A. Two categories are distinguished: targeted offense, that applies whether the message is offensive towards an individual, group, or others; and untargeted, that applies whether the message is offensive but does not contain any specific target.

- Sub-task C: Offense target identification: The task consists in identifying the type of target of an offensive message, such as an individual, a group, or any other type not fitting into the first two categories (e.g., an organization, a situation, an event, or an issue).

Sub-task A is proposed for all languages, while sub-tasks B and C are proposed for English only. Manually annotated training data are available for all languages. No language has an official development dataset. English has a special place in this edition: the organizers provided only automatically annotated material, i.e., *silver* data, together with the manually training and test data from the 2019 edition. The availability of silver data calls for innovative ways for using such data, such as a full-fledged re-training an existing pre-trained language model (as we do with mBERT) rather than directly employing them in a supervised system.

2 Related Work

Previous work on offensive language detection and related phenomena (i.e. abusive language, hate speech, cyberbullying) has seen the deployment of different system architectures with varying levels of performances. Ideally, we can observe three major waves of systems: (i.) discrete linear models (Waseem and Hovy, 2016; Karan and Šnajder, 2018); (ii.) neural networks (Cimino et al., 2018; Kshirsagar et al., 2018; Mitrović et al., 2019); and (iii.) pre-trained language models (Liu et al., 2019). Linear models (e.g. SVM, Logistic Regression, or ensemble models) are very competitive and powerful methods that have been successfully applied to identify offensive/abusive language, that in many cases outperform more complex approaches based on neural networks (Montani and Schüller, 2018). While neural networks appear to have fluctuating behaviours when applied to offensive/abusive language datasets, pre-trained language models further confirmed their predictive power.

Recently, Swamy et al. (2019) have conducted the first systematic comparison of these three families of models against four different datasets of offensive/abusive language. Feature selection and pre-processing were kept to a minimum, while they conducted fine-tuning of the hyper-parameters (i.e., sequence length, drop out, and class weights). Results confirm BERT as the best performing model. However, improvements (or decrements) across model (per dataset) are minimal. The fluctuating behavior of neural network models is further confirmed with performances being lower than those of a linear model (i.e., an SVM) in two datasets.

3 System overview

The system we propose builds on top of recent work in the use of pre-trained language models *à la* BERT (Devlin et al., 2019). We differentiate with respect to the standard fine-tuning approach by adding a retraining step. It is undisputed that BERT and BERT-like models are the new state of the art in NLP, however, these models are trained on massive amount of what could be labelled "standard" natural language data, such as news articles, Wikipedia pages, and books. None of these models is somehow "ready to be used" for Social Media data.[2] In our perspective, retraining BERT will have two beneficial effects: first, it will improve the tuning of the model towards Social Media language variety, and, second, it reduces efforts in pre-processing and cleaning of the data for fine-tuning.

[2]The only exception being AlBERTo (Polignano et al., 2019) for Italian.

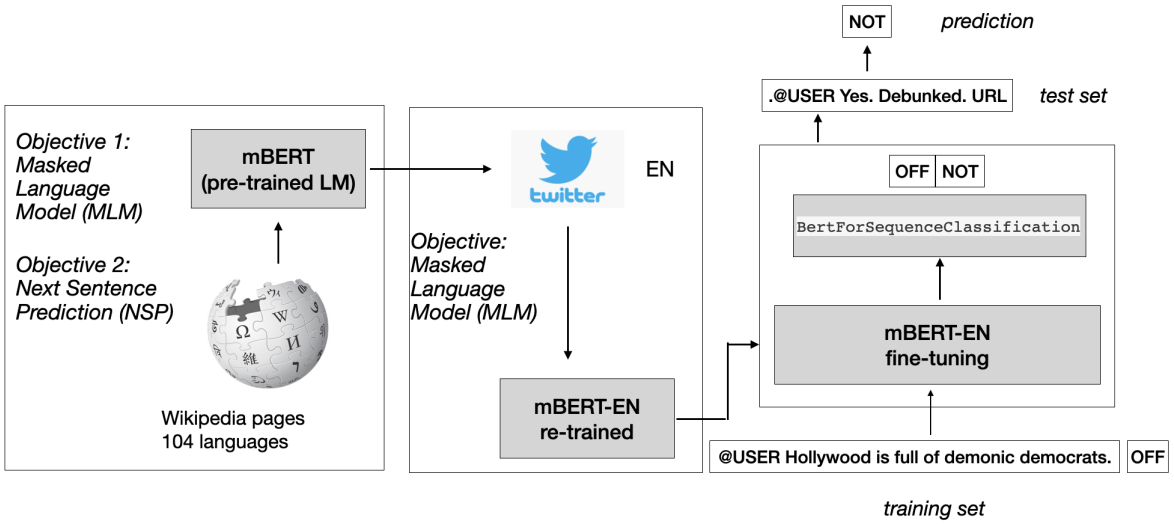

Figure 1: System illustration: the mBERT model is first re-trained using the LML objective using language-specific twitter messages, fine-tuned on the language specific training set, and then applied to classify new messages.

A further aspect we took into account is multilingualism. We aimed at developing a unifying approach that could be easily applied across the different languages. The lack of monolingual BERT models for all the languages in the task[3] guided us to select multilingual BERT (mBERT) (Devlin et al., 2019; Pires et al., 2019). mBERT consists of 12 stacked transformers, with a hidden layer size of 768 and 12 self-attention heads, like its monolingual English counterpart, $\text{BERT}_{\text{BASE}}$. mBERT is pretrained on the concatenation of monolingual Wikipedia pages of 104 languages with a shared word piece vocabulary and it does not make use of any special marker to signal the input language, nor does it have any mechanism that explicitly indicates that translation equivalent pairs should have similar representations.

Figure 1 graphically illustrates our approach. For each language, we collect potentially offensive tweets and use them to retrain the mBERT model by applying the Masked Language Model (MLM) objective. This will provide us with new "shifted" mBERT models along three dimensions: (i.) language variety (i.e. Social Media); (ii.) language (i.e. English, Danish, Turkish), and (iii.) polarity (i.e., offensive-oriented model). After retraining, the new model is fine-tuned and applied to the test data. We added a linear classifier on top of the pooled output for the [CLS] token to generate the predictions. We differentiate the general architecture only with respect to the language specific data used in the re-training and the fine-tuning steps. We developed our system for Sub-task A: Offensive language identification in three languages, namely English, Danish, and Turkish. Code, additional training data, and models are publicly available at `https://github.com/davidecolla/Offenseval2020`. The following paragraphs describes the process of collecting the additional data per language that we used to retrain mBERT.

English We created three collections of data: (i.) E1 consists of 2.5 million tokens (120,619 tweets) scraped using 736 offensive terms from the expanded version of Wiegand et al. (2018)'s lexicon; (ii.) E2 extends E1 up to 6 million tokens (283,977 tweets); (iii.) E3 extends E1 with the OffensEval 2 sub-task A data whose predicted offensive score is higher of equal to 0.6, reaching 19.5 million tokens (1,163,524 tweets).

Danish We compiled in a semi-automatic way a list of potentially offensive seed terms by combining three methods: (i.) keywords extraction using TF-IDF from the OFF messages in the training data; (ii.) the conservative portion of HurtLex v1.2 (Bassignana et al., 2018); and (iii.) a list of 140 Danish offensive terms from Wiktionary. We thus generated two collections of offensive tweets: the first, D1, contains 197k tokens (7,690 tweets). The second collection, D2, extends D1 with an additional 330k tokens obtained

[3] `https://bertlang.unibocconi.it`

using the Wiktionary list, reaching 527k tokens (20,994 tweets).

Turkish Similarly to Danish, we compiled a list of potentially offensive seed terms using the same methods: (i.) keywords from all OFF messages in the training data with TF-IDF; (ii.) the conservative portion of HurtLex; (iii.) Turkish offensive terms from Wiktionary (19 terms). We generated only one collection, T1, with 5.7 million tokens (392,674 tweets).

4 Experimental setup

We used the mBERT pre-trained model available via the huggingface Transformers library.[4] After retraining mBERT for each language, we fine-tuned the models using the training data made available by the task organizers. For English, we used the training set of OffensEval 2019. In all fine-tuned settings, we used a standard learning rate of $2e-5$, a batch size of 32, and 4 training epochs. Pre-processing steps are reported in the Appendix.

mBERT has been retrained on each of the tweet collections per language separately, generating three models for English (mBERT-E1, mBERT-E2, and mBERT-E3), two for Danish (mBERT-D1, mBERT-D2), and one for Turkish (mBERT-T1). In addition to fine-tuning the retrained models per language, we also experimented with a transfer learning approach on Danish (mBERT-D3). The choice was inspired by the close typological connection between English and Danish and the limited amount of retraining data we retrieved for Danish. We fine-tuned with the Danish training data the retrained model for English obtained with E3 (mBERT-E3). We hypothesize that mBERT-E3 could be more robust than the retrained monolingual Danish models (mBERT-D1 and mBERT-D2) because of the larger amount of retraining materials biasing mBERT for language variety and offensive content. At the same time, the typological similarity of English and Danish, and the multilingual nature of mBERT should not harm performances given the additional language specific fine-tuning step using Danish training data.

We ran an internal evaluation to verify whether the proposed system works and selected the best retrained model (at least for English and Danish). Evaluations were conducted using the OffensEval 2019 test data for English, while we split the OffensEval 2 training data for Danish and Turkish retaining 90% of the data for fine-tuning and 10% for test. On the basis of the results (see Table 3 in the Appendix for details), we selected the following systems: mBERT-E3 for English (retrained with E3 tweet collection); mBERT-D3 for Danish (transfer learning model), and mBERT-T1 for Turkish.

5 Results and Discussion

Table 1 reports the results on the blind test data. For Turkish and English our approach obtained very competitive results compared to the top ranking systems, with deltas lower than 0.05 in both cases. On the other hand, the results for the transfer learning approach in Danish are disappointing. Although we obtained very good results on the NOT class (both Precision and Recall higher than .90), transfer learning did not manage to boost the OFF class. We also evaluated the original monolingual models for Danish, mBERT-D1 and mBERT-D2. Both models obtain top-ranking macro-F1 scores (.8138 and .8195 respectively) and show a higher Precision for the OFF class when compared to the transfer learning model (.8214 and .8518 *vs.* .6285, respectively) maintaining similar Recall (.5609 for both mBERT-D1 and mBERT-D2 *vs.* .5365 for mBERT-D3). The performance on the NOT class is comparable across all models for Danish. Generally, the NOT class obtains good results across the three languages, while systems underperform against the OFF class. Table 1 also highlights a different behavior between the English model and those for the other two languages, namely a higher Recall on the OFF class. Since the main difference between the systems is the additional training data, a possible explanation for this behaviour could be a higher offensiveness load of the retraining data, that may bias the model towards the OFF class. In particular, the additional training data for our English model were collected based on higher quality lexical resources, while the Danish and Turkish data had to rely on a potentially high-coverage, but low-precision lists of lexical items.

[4] https://github.com/huggingface/transformers

Language	Model Name	Class	P	R	F1 (macro)	Δ Top Rank
EN	mBERT-E3	NOT	.9897	.8945	**.9036**	-0.018
		OFF	.7807	.9759		
DA	mBERT-D3	NOT	.9353	.9548	**.7619**	-.05
		OFF	.6285	.5365		
TR	mBERT-T1	NOT	.9002	.9342	**.7789**	-0.046
		OFF	.6967	.5935		

Table 1: Results on the official test set of OffensEval 2020.

		Predicted NOT	Predicted OFF
Actual	NOT	2511	296
	OFF	26	1054

(a) English

		Predicted NOT	Predicted OFF
Actual	NOT	275	13
	OFF	19	22

(b) Danish

		Predicted NOT	Predicted OFF
Actual	NOT	2627	185
	OFF	291	425

(c) Turkish

Table 2: Confusion matrices of the best run for each language.

Tables 2a–2c depict the confusion matrices between the predictions and the gold standard data from the task organizers. It clearly appears that the classifiers are asymmetrically biased, confirming the observation based on the scores. A qualitative error analysis on the output of the classifiers across the three languages has shown some common patterns on the reasons for the misclassifications. We have observed that False Positives ($\text{NOT}_{\text{train}} \rightarrow \text{OFF}_{\text{prediction}}$) tend to be dominated by instances containing mildly offensive terms (e.g. EN *to suck*, DA *latterlige* [ridiculous], TR *boktan* [shitty]) or terms carrying negative polarity, such as EN *ignorant*, DA *tosse* [fool]. As for the False Negatives ($\text{OFF}_{\text{train}} \rightarrow \text{NOT}_{\text{prediction}}$), we observe two trends: the first, messages contain strong offensive lexical cues that are misspelled (e.g., EN *stoopid*), or difficult to find in common lexicons of abusive terms (e.g., EN *twat*), or idiomatic expressions (e.g., TR *kapak olsun* [lit. "get a cover"]). The second concerns the presence of ambiguous words (e.g. EN *jerk*, in *@USER Wings over and it's not even a question (sweet chili & Jamaican jerk hanger)*[5] or implicitly offensive messages (DA *NED MED SVENSKEN!* [down with the Swedish][6]; TR *Şimdi sana anlatsam anlamıcan o yüzden boşver*[If I'll explain this you, you'll not understand it][7]), presence of irony (TR *@USER aşırı komikmiş kardeş ilk esprin mi* [it's too funny bro, is this your first joke?][8]), or harsh criticism (TR *Türkçe pop gibisin sesin güzel konuşmaların boş güzellik* [You sound like Turkish pop, your voice is beautiful][9]).

6 Conclusion

The approach of our system is based on the combination of re-training and fine-tuning mBERT. The re-training step has been added to bias mBERT against three aspects: language variety, language, and polarity. The bias in the classifier is sensitive to the collection of the additional training materials. Results of the systems across the three languages show that the quality of the retraining data is sensitive to the quality of the language resources and strategies used to retrieve them. On the fine-tuning step, this aspect appears to impact mainly Recall for the OFF class, as shown by the EN results compared to DA and TR.

Among the phenomena we detected as sources of noise in the classification, the degree of explicitness of the offensiveness seems to play an important role. Recent work has focused on this aspect (Kumar et al., 2020; Caselli et al., 2020) by proposing more fine-grained levels of annotations.

In future work, we will focus on the hurdles of figurative and idiomatic language usage in offensive messages, following the approach in Mladenović et al. (2017), by enriching HurtLex with multi-word expressions (MWEs) automatically extracted from corpora in multiple languages.

[5] instance ID: A2825
[6] instance ID: 1695
[7] instance ID: 32854
[8] instance ID: 38605
[9] instance ID: 43122

Acknowledgements

We want to thank Ahmet Üstun for the useful and pleasant chats during the lunch breaks which helped shaping our system, and the feedback on Turkish.

Davide Colla was spending a visiting research period thanks to an Erasmus+ Student Mobility program.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 11–16, 2020.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at evalita 2018.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium, October. Association for Computational Linguistics.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*, pages 73–77.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium, October. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression and misogyny identification in social media. In Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, France, May. European Language Resources Association (ELRA).

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 the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Jelena Mitrović, Bastian Birkeneder, and Michael Granitzer. 2019. nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 722–726, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, and Ranka Stanković. 2017. Using lexical resources for irony and sarcasm classification. In *Proceedings of the 8th Balkan Conference in Informatics*, BCI '17, New York, NY, USA. Association for Computing Machinery.

Joaquın Padilla Montani and Peter Schüller. 2018. TUWienKBS at GermEval 2018: German Abusive Tweet Detection. In *14th Conference on Natural Language Processing KONVENS*, volume 2018, pages 45–50.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Paolo Ramat and A. P. Baldry. 2011. *Linguistic Typology*. De Gruyter Mouton, Berlin, Boston.

Georg Rehm and Hans Uszkoreit, 2013. *Language Technology 2012: Current State and Opportunities*, pages 27–31. Springer Berlin Heidelberg, Berlin, Heidelberg.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, November. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

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.

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*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

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*.

Appendix

Pre-processing For the three languages, we have adopted a minimal pre-processing approach both for before the retraining and the fine-tuning steps. In particular:

- all users' mentions have been substituted with a placeholder (@USER) - only for retraining;
- all URLs have been substituted with a with a placeholder (URL) - only for retraining;
- emojis have been replaced with text (e.g. 🥺 → `:pleading_face:`) using Python `emoji` package - both for retraining and fine-tuning;
- hashtag symbol has been removed from hasthtags (e.g. kadiricinadalet → kadiricinadalet) - both for retraining and fine-tuning;
- extra blank spaces have been replaced with single spaces -both for retraining and fine-tuning.

Internal evaluation Table 3 illustrates the results of the internal evaluation to select the best system. The results support our intuitions and working hypothesis on the retraining step of mBERT.

Language	Model Name	Retrain Tokens	Class	P	R	F1 (macro)
EN	mBERT-E1	2.5M	NOT	.89	.89	.80
			OFF	.72	.70	
	mBERT-E2	6M	NOT	.89	.91	.81
			OFF	.75	.71	
	mBERT-E3	19.5M	NOT	.89	.91	**.82**
			OFF	.76	.71	
DA	mBERT-D1	197K	NOT	.91	.97	.71
			OFF	.65	.38	
	mBERT-D2	527K	NOT	.91	.96	.71
			OFF	.59	.41	
	mBERT-D3	n.a.	NOT	.92	.97	**.72**
			OFF	.64	.41	
TR	mBERT-T1	5.7M	NOT	.92	.94	**.80**
			OFF	.71	.64	

Table 3: Internal evaluation for system selection.