FBK-DH at SemEval-2020 Task 12: Using Multi-channel BERT for Multilingual Offensive Language Detection

Camilla Casula[†], **Alessio Palmero Aprosio**[‡], **Stefano Menini**[‡], **Sara Tonelli**[‡]
[†]Dept. of Linguistics and Philology, Uppsala University, Sweden
[‡]Fondazione Bruno Kessler (FBK), Trento, Italy
camillacasula@gmail.com
{aprosio,menini,satonelli}@fbk.eu

Abstract

In this paper we present our submission to subtask A at SemEval 2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval2). For Danish, Turkish, Arabic and Greek, we develop an architecture based on transfer learning and relying on a two-channel BERT model, in which the English BERT and the multilingual one are combined after creating a machine-translated parallel corpus for each language in the task. For English, instead, we adopt a more standard, single-channel approach. We find that, in a multilingual scenario, with some languages having small training data, using parallel BERT models with machine translated data can give systems more stability, especially when dealing with noisy data. The fact that machine translation on social media data may not be perfect does not hurt the overall classification performance.

1 Introduction

Concealed by perceived anonymity, users can often feel comfortable expressing offensive, abusive and hateful thoughts on the Internet. Therefore, the task of identifying offensive language on social media has increasingly gained attention in recent years, since the manual identification and deletion of such messages can be very costly and time-consuming, given the ever-increasing quantity of user-generated content online. In order to foster the development of offensive speech detection systems, more and more shared tasks have been organized on the detection and identification of offensive language in the past few years, covering several languages (Zampieri et al., 2019b; Basile et al., 2019; Wiegand et al., 2018; Bosco et al., 2018). In this paper, we present our submission to SemEval 2020 task 12, Multilingual Offensive Language Identification in Social Media (OffensEval 2) (Zampieri et al., 2020). We participated in subtask A, focused on the identification of offensive messages in 5 languages: English, Danish, Turkish, Arabic and Greek.

In the SemEval 2019 edition of OffensEval, the task was focused on Offensive Language Identification in English (Zampieri et al., 2019a). The vast majority of the teams who obtained the highest macro-F1 scores in subtask A (focused on the binary identification of offensive language) used systems based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Since the 2020 edition of the task covers five languages, we aim at improving the performance of a multilingual pre-trained BERT system by using transfer learning. This is achieved by artificially creating more data for fine-tuning two parallel pre-trained BERT models by translating existing corpora in a different language and feeding the parallel data to a multiple input model.

2 Related Work

The increased interest in the detection of offensive language on social media has entailed an increase in shared tasks on the topic (Wiegand et al., 2018; Bosco et al., 2018; Zampieri et al., 2019b), as well as more data being annotated for this task (Waseem and Hovy, 2016; Davidson et al., 2017; Golbeck et

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1539–1545
Barcelona, Spain (Online), December 12, 2020.

al., 2017; Founta et al., 2018). The topic of offensive language detection has been explored in relation to different subtasks, such as the detection of cyberbullying, hate speech, abusive language, aggression, and more (Schmidt and Wiegand, 2017; Sprugnoli et al., 2018; Zampieri et al., 2019b). In general, the methods used for identifying offensive language follow supervised learning approaches. While support vector machines perform well on the task (Schmidt and Wiegand, 2017), deep learning approaches are becoming increasingly popular (Corazza et al., 2020), in particular transformer-based ones (Zampieri et al., 2019b). More specifically, in the 2019 instance of the OffensEval task, 7 out of the 10 best performing systems used BERT-based approaches (Zampieri et al., 2019b; Devlin et al., 2019).

BERT models are typically pre-trained on large unannotated corpora. These models can then easily be fine-tuned on task-specific data, and can achieve state-of-the-art performances on a number of NLP tasks (Devlin et al., 2019). Devlin et al. (2019) provide pre-trained BERT models in English, Chinese, and a multilingual model pre-trained on the 100 largest Wikipedias[1].

Sohn and Lee (2019) successfully explore transfer learning for hate speech detection by fine-tuning a three-channel BERT model, which is fed a tri-parallel corpus consisting of data in a language and its automatic translations into English and Chinese. The data in each language is used to fine-tune the relative BERT pre-trained model and, finally, the hidden states are added together through weighted sum. This approach proves successful in improving or stabilizing the performance of systems for the identification of hate speech in Italian, German, and Spanish tweets (Sohn and Lee, 2019). We therefore build our work upon this framework using multi-channel BERT and machine-translated data.

3 Data

The data we used for participating in this task consists of the annotated datasets provided by the OffensEval task organizers and additional corpora that we use to increase the size of the training sets.

OffensEval data The organizers of the OffensEval 2 task provided the participants with a training set for each language the task was focused on: English, Danish (Sigurbergsson and Derczynski, 2020), Turkish (Çöltekin, 2020), Arabic (Mubarak et al., 2020), and Greek (Pitenis et al., 2020). The annotation guidelines and labels for all datasets are those of the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a). OLID consists of 14,100 English Tweets, annotated on three levels. The labels OFF and NOT are used to indicate whether a tweet contains offensive content, constituting the first level of annotation, used in task A.

Language	Size	% Offensive
Danish	3 k	13%
Turkish	31 k	20%
Arabic	8 k	20%
Greek	8.7 k	29%
English (OLID)	14.1k	33%
English (silver)	37.5k	31%

Table 1: Datasets provided by OffensEval organizers

An additional corpus for English was provided as "silver data" (Rosenthal et al., 2020). This corpus includes 9 million tweets automatically annotated for hate speech. The annotations consist of the average of the confidences predicted by several supervised models for a tweet to belong to the OFF class, as well as the mean standard deviation of the predicted confidences from the given average. Out of this corpus, we extracted the tweets with scores below 0.11 to be representative of the NOT class, and tweets with scores above 0.89 to represent the OFF class, amounting to 37.5k tweets overall.

The datasets are heterogeneous both in terms of size and percentage of tweets labeled as offensive. An overview of their size and distribution is provided in Table 1.

Additional data We retrieve additional training data in English by merging five widely used datasets for hate speech / abusive language detection, namely the ones from Waseem (2016), Waseem and Hovy (2016), Davidson et al. (2017), Golbeck et al. (2017) and Founta et al. (2018), after mapping their annotation to the OFF/NOT classes. We then extract around 30% of the tweets (27k messages) to extend the task training data. We choose not to use all the merged data because the tweets are then automatically translated and we want to limit the impact of wrong translations on classification performance.

[1] github.com/google-research/bert

Final datasets For each language other than English, we evaluate three different datasets: (1) Language-specific OffensEval corpus only (OE), (2) OE + OLID automatically translated into the language of interest (i.e. OLID-DA, OLID-EL, OLID-AR and OLID-TR), (3) OE + translated OLID + additional 27k tweets from merged datasets, automatically translated into the language of interest.

All translations have been performed using Google APIs. For the classification of English tweets, we used the OLID tweets and the filtered silver dataset as training data (last two rows in Table 1).

4 Classification Framework

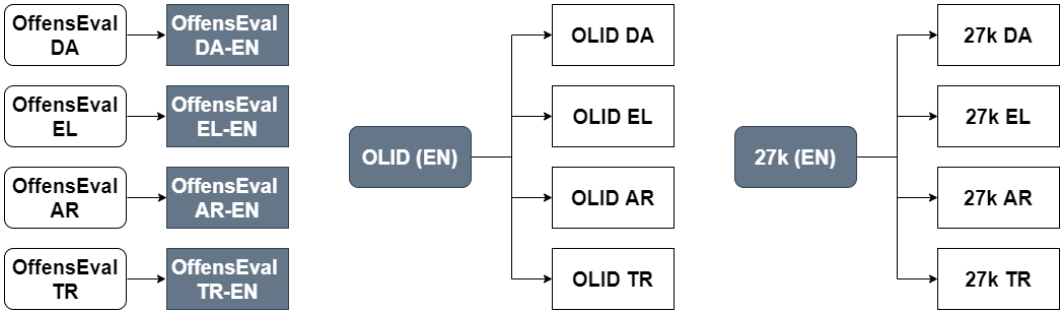

Figure 1: A summary of the dataset translations we performed. English data is on dark background.

4.1 Danish, Turkish, Arabic, and Greek: Multi-channel BERT

Baselines For Danish, Turkish, Arabic, and Greek, our baselines consist of a single-channel BERT multilingual model fine-tuned on the available data for each language. A subset of this data (20%) is used as development set.

Pre-processing In all the data we use, user mentions are already replaced with `@USER` and URLs are replaced with `URL` tags.

Approach We build a multi-channel BERT system inspired by that of Sohn and Lee (2019), in which we fine-tune both a pre-trained English BERT base cased model and a pre-trained multilingual BERT cased model in the same architecture. Our methodology relies on two main steps. First, we create multiple parallel offensive language detection corpora by translating the corpus available for each of the four languages (the OffensEval 2020 corpora) into English and the corpora available in English (OLID and 27k) into each of the four languages. A summary of all the translation processes can be found in Figure 1. With this process, we significantly extend the training data available for each language, although some tweets may not be correctly translated.

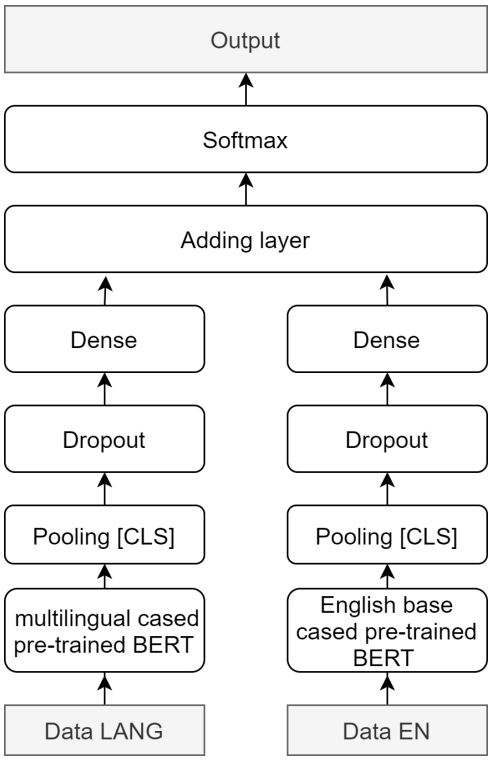

Figure 2: The multi-channel model architecture. *LANG* is used to refer to any of the four languages in the multilingual task.

The second step consists in fine-tuning two BERT models in parallel: multilingual BERT cased and English BERT base cased. Our model takes as input two parallel datasets, one in the target language (LANG) and one in English.

We initialize the model weights for each model using the parameters from the two pre-trained BERT models. Then, the hidden representation of the first token (`[CLS]`) of each sequence is pooled out through a pooling layer. We then add a dropout layer (with 0.1 dropout rate) and a dense layer (768 units, ReLu activation and L2 regularization), before adding the hidden states of each BERT model together.

Finally, we add a softmax layer for classification. For all our experiments, both on our baseline and on the multi-channel model, we use batch size 32, sparse categorical cross-entropy loss, and Adam optimizer with learning rate 2e-5, similarly to Sohn and Lee (2019). A summary of our architecture can be seen in Figure 2. The code used for our system is available at `github.com/ca-milla/multi-channel-bert`. In order to choose our systems for the submission, a random sample of the training data in each configuration was retained to be used as development set.

4.2 English: Standard monolingual BERT

Our experiments on English are performed on OLID (after we leave out 1k randomly selected tweets as development set) and a portion of the distantly annotated corpus of 9 million English tweets provided by the task organizers (Zampieri et al., 2020), which was selected as described in Section 3. The total training data available after merging the two datasets and removing duplicates was 48,581 tweets, of which 33,172 tweets labeled as NOT and 15,409 were labeled as OFF.

Baseline Our baseline for English consists of a pre-trained BERT English base cased model trained on the same data used for the system submission, but without any pre-processing.

Pre-processing The datasets provided by the organizers included `<user>` tags for user mentions and `<url>` tags for URLs. We normalize the data and split hashtags and elongated words using *ekphrasis*[2], a tool made specifically for social media data normalization.

Approach Our architecture for the English submission consists of a pre-trained BERT English base cased model, fine tuned on the corpus containing 48,581 tweets previously described. After initializing the model weights using the pre-trained BERT weights, the hidden representation of the first token (`[cls]`) is pooled and fed into a dropout layer with dropout rate 0.2. Two dense layers are then added (sizes 768 and 128) and, finally, a softmax layer is used for classification. The hyperparameters used are the same used for the multilingual systems.

5 Results and Discussion

	Danish		Turkish		Arabic		Greek	
	Baseline	Multi-channel BERT	Baseline	Multi-channel BERT	Baseline	Multi-channel BERT	Baseline	Multi-channel BERT
OE	0.41	0.41	0.43	0.43	0.44	0.44	0.70	0.42
OE + OLID	0.41	0.74	0.43	0.43	0.63	0.77	0.69	0.77
OE + OLID + 27k	0.70	**0.78**	0.43	**0.74**	0.60	**0.79**	0.73	**0.79**

Table 2: Macro F1 scores obtained by our baseline and system on development data.

5.1 Multi-channel BERT

The results obtained by our baseline and our multi-channel BERT system on the development set are reported in Table 2. In general, multi-channel BERT was the best performing system, especially when paired with large quantities of data. We therefore used multi-channel BERT trained on all available data for our submissions on Danish, Turkish, Arabic, and Greek. Our main finding is that while a multi-channel setting can offer little improvement when dealing with homogeneous language-specific data (as is the case for the OffensEval corpora), it can improve systems when only small training data are available by adding machine-translated data from different datasets. This could offer insight for research on low-resource languages in the field of offensive language detection.

Danish We find that the multi-channel BERT architecture, featuring both BERT multi pre-trained and BERT English base pre-trained can perform better than a system using BERT multilingual alone. This is especially the case when we feed the system more data, since both models assign our development set the most frequent label when fine-tuned on OffensEval data alone. The best performing combination

[2] `github.com/cbaziotis/ekphrasis`

(BERT multi-channel fine-tuned on all data available) was chosen for our submission. Our macro F1 score on the OffensEval test set was **0.777**, and among our submitted runs it was the best ranked one, scoring 4th in the ranking for Danish (the best performing system achieved 0.812 F1). The integration of machine-translated data probably yields good results because English and Danish are typologically similar, being both Germanic languages, which could positively affect the translation quality.

Turkish Turkish was the language with the largest OffensEval corpus. However, both systems assigned the most frequent label when tested on our dev set both when fine-tuned on Turkish OffensEval data alone and when fine-tuned on OLID. Our best performing configuration is multi-channel BERT trained on all data available. This is the configuration we used for our OffensEval submission. The macro F1 score we obtained on the OffensEval Turkish test set is **0.62**, almost 0.20 F1 lower than the top-ranked one.

Arabic As with Danish and Greek, both the baseline and multi-channel BERT assign the most frequent label to instances in our development set when fine-tuned on OffensEval data alone. Our best performing system, which was used for our submission, is BERT multi-channel trained on all available data. On the OffensEval Arabic test set, our multi-channel system achieves a macro F1 score of **0.46**. The contrast between the score obtained on our development set and the test set of the task is still an open issue which we would like to explore in the future, also checking wrong classifications by hand.

Greek When fine-tuned on OffensEval data only, the pre-trained BERT multi model performs much better than the multi-channel BERT. However, when the systems are fine-tuned on larger quantities of data, i.e. in the OE+OLID and OE+OLID+27k configurations, the multi-channel BERT architecture achieves higher macro F1 scores. Our best performing system is again BERT multi-channel fine-tuned on all data available. This is also the system we used for our submission. Our macro F1 score on the OffensEval Greek test data is **0.77**, i.e. 0.08 F1 lower than the best performing system.

5.2 English

The F1 scores obtained on our English development set, consisting of a randomly selected subset of OLID, can be found in Table 3. We notice improvements in the performance of the system when fine-tuning the pre-trained BERT base model on normalized data, reducing data sparsity problems, mostly due to elongated words and hashtags. Our system achieved a macro F1 score of 0.903 on the task test data. Most of the runs submitted to the task yield F1 > 0.90, showing that when large amounts of training data are available, the architecture chosen for classification is not very relevant, in that also simple ones like ours obtain very good results.

	OE + OLID
Baseline (BERT EN - no pre-processing)	0.76
Pre-processing + BERT EN	**0.79**

Table 3: Macro F1 scores obtained on our English development set.

6 Conclusion

In this paper, we present our system participating in the SemEval 2020 task 12. Our goal was not to develop different systems optimized for each language of the task, but rather to implement a framework based on two-channel BERT and machine translated data that could be applied to Danish, Turkish, Arabic and Greek without many changes. We show that this approach yields good results, always outperforming single-channel BERT, especially when few training data are available, since the system takes advantage of larger machine translated datasets in English. The most promising results using multi-channel BERT are obtained on Danish, and we will explore in the future whether it depends on the small size of the training data or on the similarities between English and Danish, which may positively affect translation quality.

While our approach is inspired by Sohn and Lee (2019), there are some differences both in the architecture and in the results: our framework is simpler, in that we implement a two-channel BERT, while they rely on three channels using also machine-translated texts from Chinese. This makes our framework less computationally-intensive. As regards the results, we observe that in our experiments multi-channel BERT is generally beneficial, while in Sohn and Lee (2019) results on Italian and German are better when using a single channel.

In the future, we plan to perform an error analysis to understand why the run we submitted for Arabic performed remarkably worse than the results obtained on the development set. We will also evaluate the classifier performance when using only machine-translated data, in a zero-shot learning setting.

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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 366–370, Marseille, France. European Language Resources Association.

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*, 20(2):10:1–10:22.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

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*, pages 4171–4186.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

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*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France, may. European Language Resources Association.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

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*, pages 3498–3508, Marseille, France, may. European Language Resources Association.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. 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, San Diego, California, June. Association for Computational Linguistics.

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, Austin, Texas, November. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1 – 10, Vienna, Austria. Austrian Academy of Sciences.

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.

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*. Association for Computational Linguistics.