SINAI at SemEval-2020 Task 12: Offensive language identification exploring transfer learning models

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González,
L. Alfonso Ureña-López, M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, laurena, maite}@ujaen.es

Abstract

This paper describes the participation of SINAI team at Task 12: OffensEval 2: Multilingual Offensive Language Identification in Social Media. In particular, the participation in Sub-task A in English which consists of identifying tweets as offensive or not offensive. We preprocess the dataset according to the language characteristics used on social media. Then, we select a small set from the training set provided by the organizers and fine-tune different Transformer-based models in order to test their effectiveness. Our team ranks 20th out of 85 participants in Subtask-A using the XLNet model.

1 Introduction

In recent years the way of sharing information has suffered a change by the worldwide proliferation of use of social media. However, the freedom of this use often leads many users to communicate using offensive language.

Offensive language is defined as the text which uses hurtful, derogatory or obscene terms made by one person to another person (Wiegand et al., 2019). One of the strategies used to deal with offensive behavior on social media is to monitor or report this type of comment by users. However, this strategy is not entirely feasible due to the huge amount of data generated daily by users. Therefore, it is necessary to develop automatic systems capable of identifying this type of content on the Web.

In this paper, we present the systems we developed as part of our participation in SemEval-2020 Task 12: OffensEval 2 - Multilingual Offensive Language Identification in Social Media (Zampieri et al., 2020). In particular, the participation in Sub-task A in English which consists of identifying tweets as offensive or not offensive.

The rest of the paper is structured as follows. In Section 2 some previous related studies are introduced. In Section 3, we describe the pre-trained models we have studied to address the task. In Section 4 we explain the data used in our methods, we present the details of the proposed systems and we discuss the analysis and evaluation results for our system. Finally, we conclude in Section 5 with remarks.

2 Background

In recent years, the textual analysis of offensive detection has attracted the attention of several researchers in the field of Natural Language Processing (NLP) leading to various academic events and shared tasks in different languages such as the first, second and third editions of the Workshop on Abusive Language (Roberts et al., 2019); the first and second edition of GermEval Shared Task on the Identification of Offensive Language (Struß et al., 2019); the shared task on aggression identification included in the First Workshop on Trolling, Aggression and Cyberbullying (Kumar et al., 2018); the MEX-A3T track at IberLEF 2019 on Authorship and Aggressiveness Analysis (Aragón et al., 2019); the PolEval 2019 shared Task 6 on Automatic Cyberbullying Detection in Polish Twitter (Ptaszynski et al., 2019); the first edition of the HASOC track at FIRE 2019 on HS and Offensive Content Identification in Indo-European

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1622–1627
Barcelona, Spain (Online), December 12, 2020.

Languages (Mandl et al., 2019); and finally the subtask 6 on Identifying and categorizing offensive language in social media (OffensEval) (Zampieri et al., 2019).

These competitions have led to very promising systems to combat offensive language using different techniques. For instance, in OffensEval last year's edition participants generally used systems based on deep learning and traditional machine learning. The top teams used ensembles and state-of-the-art deep learning models such as pre-trained models based on transfer learning. Transfer learning is a recently emerging technique to enhance learning in a new task through the transfer of knowledge from a related task that has already been learned (Torrey and Shavlik, 2010). It is widely used to develop models to solve problems where the availability of large data is limited in order to train, validate, and evaluate the systems. For instance, the top-performing team (Liu et al., 2019a) in Sub-task A used different systems (a linear model, an LSTM and BERT) obtaining the best result with BERT-base-uncased model. Mahata et al. (2019) ranked sixth using an ensemble of CNN and BLSTM-BGRU together with Twitter word2vec embeddings and token/hashtag normalization. Nikolov and Radivchev (2019) trained a variety of models such as CNN, MLP, SVM or RNN combined the best of them in ensembles, but this team proved again that BERT is a powerful tool for offensive text classification as it reached the best result.

Over the last two years, other Transformer-based models based on BERT such as RoBERTa, XLNeT or DistilBERT have been presented to improve BERT on either its prediction metrics or computational speed. In this paper, we focus on explore the effectiveness of some of these models for offensive language detection adapting a small version of the training set provided by the organizers.

3 System overview

In this section, we describe the pre-trained models we have explored to address the task of offensive language identification. They are based on Transformer which is an encoder-decoder architecture based on the attention mechanism (Vaswani et al., 2017).

XLNet (Yang et al., 2019) is a generalized autoregressive pre-training method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. This allows the model to learn bidirectional relationships and consequently better handles dependencies and relations between words. XLNet outperforms BERT on 20 tasks, including question answering, natural language inference, sentiment analysis, and document ranking.

RoBERTa (Robustly optimized BERT) (Liu et al., 2019b) approach arises to address some of BERT's limitations improving training procedure since it was significantly undertrained. In order to do that, the following modifications were carried out: training the model longer, with bigger batches, over more data; removing the next sentence prediction objective from BERT's pre-training; training on longer sequences; and introduces dynamic masking.

DistilBERT (Sanh et al., 2019) is a general-purpose pre-trained version of BERT, 40% smaller, 60% faster, that retains 97% of the language understanding capabilities. DistilBERT uses a technique called distillation, which is a compression technique in which a small model is trained to reproduce the behavior of a larger model (Buciluǎ et al., 2006), in this case the large model is BERT. The idea is that once a large neural network has been trained, its full output distributions can be approximated using a smaller network. DistilBERT was trained on very large batches leveraging gradient accumulation, with dynamic masking and removed the next sentence prediction objective.

4 Experimental setup

4.1 Dataset and training

To run our experiments, we use the English dataset provided by the organizers in SemEval 2020 Task 12: OffensEval 2 - Multilingual Offensive Language Identification in Social Media (Rosenthal et al., 2020). In particular, we use the dataset for Sub-task A which consists of identifying a tweet as offensive (OFF) or non-offensive (NOT). The dataset contains tweets with several fields including an identifier (ID), the text of the tweet (TWEET), the average of the confidences predicted by several supervised models for a specific instance to belong to the positive class (AVG_CONF) and the average of the confidences predicted by several supervised models for a specific instance to belong to the positive class (CONF_STD).

In first place, we preprocess the corpus of tweets applying the following steps: the URLs and mentions users are removed and the hashtags are unpacked using the ekphrasis tool[1].

Due to the large number of tweets in the original dataset (about 9 million tweets) to train our system, we consider selecting a smaller number of tweets in order to test the effectiveness of pre-trained models fine-tune them on a small dataset. In order to build our training small dataset we follow these steps:

1. We discard tweets that are not clearly defined as OFF or NOT by using the AVG_CONF and CONF_STD fields of each tweet. We define a Threshold Level (TL) of 0.5. In order to label the tweet as OFF or NOT, we perform the following equation:

$$f(x) = \begin{cases} OFF & if(AVG\_CONF_{tweet} \pm CONF\_STD_{tweet}) \geq TL \\ NOT & if(AVG\_CONF_{tweet} \pm CONF\_STD_{tweet}) < TL \\ & \\ DISCARD & if((AVG\_CONF_{tweet} + CONF\_STD_{tweet}) \geq TL) \\ & and(AVG\_CONF_{tweet} - CONF\_STD_{tweet}) < TL) \end{cases} \quad (1)$$

   A total of 872,069 and 6,346,677 tweets are classified as OFF and NOT respectively, and 1,870,394 tweets are discarded.

2. For the purpose of obtaining a smaller set while having the same proportion of tweets as the dataset selected in the previous step, we determine four groups depending on their AVG_CONF. From each group we choose 1 tweet out of 100, thus reducing to 1% the number of tweets used for the training phase. This process can be seen in Table 1.

Range (AVG_CONF)	Tweets by group	Tweets selected
1 to 0.75	542,708	**5,427**
0.75 to 0.5	329,361	**3,293**
0.5 to 0.25	1,773,015	**17,730**
0.25 to 0	4,573,662	**45,736**
Total tweets	7,218,746	**72,186**

Table 1: Number of tweets selected based on their AVG_CONF for the training set.

Finally, our new training dataset generated from the English dataset provided by the organizers is composed of 72,186 tweets (8,720 labeled as OFF and 63,466 labeled as NOT).

4.2 Transfer learning fine-tuning

Transfer learning has emerged as a highly popular technique in developing deep learning models. In this technique, the neural network is trained in two stages: 1) pre-training, where the network is generally trained on a large dataset representing a wide diversity of labels or categories; and 2) fine-tuning, where the pre-trained network is further trained on the specific task, which may have fewer labeled examples than the pre-training dataset. The first step helps the network learn general features that can be reused on the target task.

In this task we address the identification of offensive language in tweets using different transfer learning models. Therefore, it is essential to analyze the contextual information extracted from the pre-trained layers of these models and then fine-tune it using a labeled dataset. By fine-tuning we update weights using the small training dataset generated that is new to an already trained model.

For the purposes of fine-tuning, it is recommended choosing from the following values: batch size, learning rate, max sequence and number of epoch. We empirically set the number of epochs to 3 and the learning rate to 2e-5. For the optimizer, we leverage the *adam* optimizer which performs well for NLP data and for BERT models in particular. We set the batch size to be 32 for BERT and DistilBERT and

[1] https://github.com/cbaziotis/ekphrasis

8 for XLNet and RoBERTa. The max sequence length was established to 80 in all the experiments and they were run on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

4.3 Results

In this section we present the results obtained by the systems we have explored. In order to evaluate them we use the official competition metric macro-averaged F1-score. In addition, we show another usual metrics employed in classification tasks including Precision (P) and Recall (R).

The results of our participation in the Subtask-A in English during the evaluation phase can be seen in Table 2. In particular, we list the performance of our three selected models which are RoBERTa, DistilBERT and XLNet. As it can be seen, the results achieved by the three models are nearly the same, RoBERTa and DistilBERT report the same value of macro-F1 score (0.909) and the best performance is achieved by XLNet with a macro-F1 score of 0.911. It should be noted that the OFF class gets the lowest score in the three models due to the imbalance distribution of the dataset as the precision score reports the lowest result in this class. Our results show the success of the pre-trained models we chose in order to solve the task of offensive language identification in tweets using a small training set adapted from the original provided by the organizers.

System	NOT			OFF			Macro-Avg		
	P	R	F1	P	R	F1	P	R	F1
RoBERTa	0.996	0.895	0.943	0.785	0.991	0.876	0.890	0.943	0.909
DistilBERT	0.994	0.897	0.943	0.787	0.986	0.876	0.891	0.942	0.909
XLNet	0.996	0.897	**0.944**	**0.787**	**0.992**	**0.878**	**0.892**	**0.944**	**0.911**

Table 2: Systems test results per class in Subtask-A of OffensEval task.

In Table 3 we can observe our official rank in the competition. We are ranked 20 out of 85 participating teams obtaining a Macro-F1 score of 0.911 with the XLNet model. It is noteworthy that the best result obtained by the best participant in Sub-task A differs slightly from our result (1,1%). We also observe that the number of participants in the sub-task is high (84) which shows the importance and interest of the NLP community in helping to solve this type of tasks.

User name (ranking)	Macro-F1
ltuhh2020 (1)	0.922
SpurthiAH (10)	0.914
fmplaza (20)	**0.911**
hussam (50)	0.907
zampieri (84)	0.419

Table 3: System Results per participating team in Subtask-A of OffensEval task.

5 Conclusions

The use of offensive language is a growing concern on social media due to the serious consequences it can have for users. To help combat this problem, automatic systems based on NLP can be very effective. For instance, transfer learning techniques have been proven to be especially effective for several tasks including classification. The nature of informal language which involves the use of abbreviations, dialect words, colloquial expressions or emojis, together with the limited length of the publications in case of Twitter, make the task even more challenging. In this paper, we show that the results obtained by the pre-trained models we have explored are promising using an adequate preprocessing in tweets. Generally, it is not easy to have labeled data because of the high cost and effort involved. For this reason, transfer learning techniques are a valuable resource for carrying out these types of tasks.

Acknowledgements

This work has been partially supported by LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER).

References

Mario Ezra Aragón, Miguel Á Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018. Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

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

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*.

Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Ratn Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 683–690, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

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.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval 2019 Workshop*, page 89.

Sarah T Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem. 2019. Proceedings of the third workshop on abusive language online. In *Proceedings of the Third Workshop on Abusive Language Online*.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language.

Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global.

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.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

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, 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*.