UTFPR at SemEval 2020 Task 12: Identifying Offensive Tweets with Lightweight Ensembles

Marcos Aurélio Hermogenes Boriola, Gustavo Henrique Paetzold
Universidade Tecnológica Federal do Paraná, Toledo-PR, Brazil
marcosboriola@alunos.utfpr.edu.br, ghpaetzold@utfpr.edu.br

Abstract

Offensive language is a common issue on social media platforms nowadays. In an effort to address this issue, the SemEval 2020 event held the OffensEval 2020 shared task where the participants were challenged to develop systems that identify and classify offensive language in tweets. In this paper, we present a system that uses an Ensemble model stacking a BOW model and a CNN model that led us to place 29th in the ranking for English sub-task A.

1 Introduction

Social media websites are widespread on the internet and they are used to facilitate the communication between people. Their usage is growing more and more (Perrin, 2015). People use these means of communication to express their opinions, to share something they like, to keep in touch with their friends and relatives, and also to get to know someone new. However, some people use social media to commit attacks against others, consequently sharing offensive content on theses platforms (Chetty and Alathur, 2018). In an effort to help identify offensive language on social media posts, we rely on Natural Language Processing (NLP) techniques, since this is an area of Computer Science and Linguists that cares about the interaction of computers and humans through natural language. And this kind of task in NLP is called Binary Text Classification, where, based on textual features, a system estimates whether or not a sentence has offensive language.

There are many types of approaches to solving a Text Classification problem, such as statistical models (e.g., Naive Bayes, SVM), machine learning models (e.g., gradient boosted decision trees), deep learning models (e.g., LSTM, RNN, CNN, BERT), and other methods as well. Last year, OffensEval 2019 shared task participants created many approaches to solve the issue of offensive language identification, where 70% of them used deep learning models as a solution to this task. Analyzing only the methods used by the top-10 teams, seven of them used BERT to solve the task (Zampieri et al., 2019). Although BERT has managed to achieve state-of-the-art results in many NLP tasks (Devlin et al., 2018), it is very computationally expensive to train compared to most machine learning models (Peng, 2019). Due to its complexity and robustness, the BERT model requires many GPUs for its training and this model is not very suitable for low resource languages.

In an attempt to solve this problem, we developed a more lightweight approach using an ensemble model stacking a Bag-of-words (BOW) model and a Convolutional Neural Network (CNN). The BOW model can be easily applied to even low resource languages and the CNN model is considerably less costly to train than the BERT model. We claim that's a lightweight model due to a runtime of 11 hours in a freely available virtual machine (1-core/2-threads Xeon CPU + Tesla K80 GPU + 25GB of RAM), which is way less costly than the models mentioned before. In the following sections, we describe the task (Section 2), our system (Section 3), and its performance on the shared task (Section 4).

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2232–2236
Barcelona, Spain (Online), December 12, 2020.

2 Task Summary

The OffensEval shared task is part of the SemEval 2020 workshop (Zampieri et al., 2020) and consists of developing a classification system for each of the sub-tasks described below:

- **Task A:** Categorization of a tweet as offensive or not.
- **Task B:** Classification of an offensive tweet as targeted or not to someone.
- **Task C:** Identification of the target of the offensive tweet (individual, group, or other).

To help the participants create a solution for the shared task, datasets in five different languages were provided (Arabic (Mubarak et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), English (Rosenthal et al., 2020), Greek (Pitenis et al., 2020), and Turkish (Çöltekin, 2020)). However, only the English dataset ran all three sub-tasks, the other languages ran only sub-task A. Our team focused on the English dataset, so only this dataset will be described ahead. For sub-task A, the organizers provided a dataset containing 9,087,118 instances; for sub-task B, a dataset containing 188,973 instances; and for sub-task C, a dataset containing 188,973 instances.

Our team focused on developing a system based on the English dataset but only for solving the sub-task A. For this sub-task, a total of 9,087,118 instances were released to be used as training and development set and 3,887 instances as test set.

Analyzing the training/dev sets, we found they are tabular datasets separated into 4 columns: The first one is the id, the second one is the sentence, the third one is the average, and the last one is the standard deviation (see Table 1). The values on the third and fourth columns are the results estimated by a weakly supervised model trained by the task organizers (Rosenthal et al., 2020). The range of values on the average column is from 0 to 1, where a value near 0 represents a sentence less offensive and a value near 1 represents a sentence more offensive.

id	text	average	std
1159533793511972865	@USER Not really his job tho.	0.2950074258748575	0.146550464816645
1159528564925984768	everyone talks shit in LA	0.8600965119695152	0.15907768688355775
1159533703758061570	@USER His ass need to stay up	0.8333493190791922	0.14069975111613903
1159533780945838081	@USER his bouffant tail is amazing	0.18088214387350932	0.14578659894547777

Table 1: Sentences extracted from Training/Development dataset.

3 The UTFPR System

3.1 Pre-processing

As the objective of this task is to classify a sentence as offensive (OFF) or not offensive (NOT) but the data was only labeled with numerical offensive scores, we set a threshold of 0.5 to determine what is offensive and what is not offensive. Every sentence that has an average value under 0.5 was judged as not offensive (NOT), thus, every sentence with an average value equal or above 0.5 was judged as offensive (OFF). After doing that, this new dataset was analyzed and we could observe that the dataset was imbalanced, since we got considerably more NOT labels than OFF labels. In order to get it balanced and get better performance on our system, we used an under-sampling method to get equal counts for each label. The under-sampling method randomly selected the sentences from the NOT label so that we obtained a completely balanced dataset (see Table 2 for more details).

Class	Before under-sampling	After under-sampling
NOT	$7,637,449$ (84.05%)	$1,449,669$
OFF	$1,449,669$ (15.95%)	$1,449,669$
All	$9,087,118$	$2,899,338$

Table 2: Number of instances in the training/development dataset before and after under-sampling.

After balancing the dataset, no further processing was done on the training/dev dataset. We also decided to create a pseudo test set using the dataset provided in OffensEval 2019 so that we could use the OffensEval 2020 dev set for validation during training. The pseudo test set is a combination of the training and test sets provided for OffensEval 2019 sub-task A. Table 3 contains details about the quantity of instances, as well as OFF and NOT labels in our pseudo test set.

	OffensEval 2019 Training set	**OffensEval 2019 Test set**	**Pseudo test set**
NOT	8,840	620	9,460
OFF	4,400	240	4,640
Total	13,240	860	14,100

Table 3: Number of instances in the OffensEval 2019 dataset and our pseudo test set.

3.2 Methodology

BOW: The bag-of-words model is a method to represent texts numerically. Given a set of texts, this model counts how many times a word appears in each text. Thus, using the data provided for the task, the whole training set is the corpus used to get the vocabulary from and each sentence is further vectorized in an array where each column represents how many times that word appeared in the sentence.

CNN: The Convolutional Neural Network (CNN) was first presented as an approach to object recognition in 1999 (LeCun et al., 1999). Since then, this kind of neural network has been applied as a solution to many other problems, even in Natural Language Processing tasks, and one of them is the Text Classification (Bhandare et al., 2016). A CNN model can detect variations of patterns in the input data, which promotes good feature extraction even from dirty data (Kim, 2014). Looking for a robust and computationally efficient model, we choose this approach to address this task. This choice was backed too by the good results obtained by other teams that used CNN models at OffensEval 2019 (Zampieri et al., 2019).

spaCy TextCategorizer: The TextCategorizer[1] is a spaCy[2] model that, as its name suggests, is used for text classification and it is the model used for the system herein described. Inside of TextCategorizer, there are four parameters: vocab, which is a Vocab object that is the vocabulary that will be used in the model; model, the language model used which, if not provided, will be created based on the data provided; exclusive_classes (True or False), which decides whether the classes provided will be mutually exclusive; and architecture ("ensemble", "simple_cnn", and "bow"), which is the architecture utilized by the classifier. Below are some relevant parameters we used for our model:

- **spaCy version:** 2.2.3
- **Parameters:**
  - **exclusive_classes:** False
  - **model:** en_core_web_lg[3] v2.2.5
  - **architecture:** ensemble
- **Number of iterations:** 10.

Training: We used the OffensEval 2020 training set to train the models, the OffensEval 2020 dev set to validate the models during training, and chose as our final submitted model the one with the highest score over the pseudo test set.

4 Performance on Shared Task

The official evaluation metric used for the OffensEval 2020 shared task is the macro-averaged F1-score. In Table 4 are the results obtained by the UTFPR system on the pseudo test set and the official OffensEval

[1]https://spacy.io/api/textcategorizer
[2]https://spacy.io
[3]https://spacy.io/models/en#en_core_web_lg

2020 test set. As it can be noticed, the official test set scores were better than expected compared to the pseudo test set. Analyzing the results in order to discover where our system had failed to classify some tweets, we found out that our system tends to classify as offensive ("OFF" label) tweets that contain bad words even when these words are used to emphasize something. The model misclassifies as not offensive ("NOT" label) tweets that have any level of irony.

Dataset	F-score
Pseudo test set	0.793
Official test set	0.909

Table 4: Macro-averaged F-scores of the UTFPR system on the pseudo test set and the official test set.

Table 5 shows the top 3 and bottom 3 team scores on sub-task A. The score our system obtained placed us in 29th place on the ranking, but we see a difference of less than 2% from top 1.

Rank	Team	F-score
1	UHH-LT	0.9204
2	Galileo	0.9198
3	Rouges	0.9187
29	UTFPR	0.9094
79	KarthikaS	0.6351
80	Bodensee	0.4954
81	**Majority Baseline**	0.4193
82	IRlab@IITV	0.0728

Table 5: Ranking scores.

5 Conclusions

The UTFPR system herein presented for the OffensEval sub-task A used an ensemble model that stacks a BOW and a CNN model. Despite being lightweight and easily adaptable to low-resource languages, our model performs well when compared to more sophisticated and resource-dependent systems. In this task our system got the 29th position out of a total of 82 participants staying only 2% F-score points away from first place.

Malicious users on social media platforms generally write their posts using spelling variations on bad words to bypass algorithms that check the offensive language in their texts. Because of that, in the future, we plan to use text normalization techniques (mainly on incorrectly-spelled words) in order to train more robust and reliable models for this task.

6 Acknowledgments

We gratefully acknowledge the support of the Universidade Tecnológica Federal do Paraná - Campus Toledo, which provided other important resources that were crucial in the development of this contribution.

References

Ashwin Bhandare, Maithili Bhide, Pranav Gokhale, and Rohan Chandavarkar. 2016. Applications of convolutional neural networks. *International Journal of Computer Science and Information Technologies*, 7(5):2206–2215.

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

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. 1999. Object Recognition with Gradient-Based Learning. In D. Forsyth, editor, *Feature Grouping*. Springer.

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*.

Tony Peng. 2019. The Staggering Cost of Training SOTA AI Models. *Synced. URL https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/*.

Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.

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.

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 (SemEval)*.

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*.