ConvAI at SemEval-2019 Task 6: Offensive Language Identification and Categorization with Perspective and BERT

John Pavlopoulos
Ion Androutsopoulos
Department of Informatics
Athens University of Economics
and Business, Greece
annis,ion@aueb.gr

Nithum Thain
Lucas Dixon
Jigsaw
nthain,ldixon@google.com

Abstract

This paper presents the application of two strong baseline systems for toxicity detection and evaluates their performance in identifying and categorizing offensive language in social media. **Perspective** is an API, that serves multiple machine learning models for the improvement of conversations online, as well as a toxicity detection system, trained on a wide variety of comments from platforms across the Internet. **BERT** is a recently popular language representation model, fine tuned per task and achieving state of the art performance in multiple NLP tasks. **Perspective** performed better than **BERT** in detecting toxicity, but **BERT** was much better in categorizing the offensive type. Both baselines were ranked surprisingly high in the SEMEVAL-2019 OFFENSEVAL competition, **Perspective** in detecting an offensive post (12th) and **BERT** in categorizing it (11th). The main contribution of this paper is the assessment of two strong baselines for the identification (**Perspective**) and the categorization (**BERT**) of offensive language with little or no additional training data.

1 Introduction

Offensive language detection refers to computational approaches for detecting abusive language, such as threats, insults, calumniation, discrimination, swearing (Pavlopoulos et al., 2017b), which could be targeted (at an individual or group) or not (Waseem et al., 2017). These computational approaches are often used by moderators who face an increasing volume of abusive content and would like assistance in managing it efficiently.[1]

Although offensive language detection is not a new task (Dinakar et al., 2011; Dadvar et al., 2013; Kwok and Wang, 2013; Burnap and Williams, 2015; Tulkens et al., 2016), the creation of large corpora (Wulczyn et al., 2017), along with recent advances in pre-training text representations (Devlin et al., 2018) allow for much more efficient approaches. Furthermore, while new competitions and corpora are being introduced (Zampieri et al., 2019a),[2] there is a need for strong baselines to assess the performance of more complex systems. This paper assesses two systems for the detection and categorization of offensive language, which require few or no task-specific annotated training instances.

The first baseline is a Convolutional Neural Network (CNN) for toxicity detection, trained on millions of user comments from different online publishers, which is made publicly available through the **Perspective** API.[3] This model requires no extra training or fine tuning and can be directly applied to score unseen posts. The second strong baseline is the recently popular Bidirectional Encoder Representations from Transformers (**BERT**), a pre-trained model that has been reported to achieve state of the art performance in multiple NLP tasks with limited fine-tuning on task-specific training data (Devlin et al., 2018).

Section 2 below summarizes related work and Section 3 discusses the SEMEVAL-2019 OFFENSEVAL dataset we used. In Section 4 we describe the two proposed baselines and we report experimental results in Section 5. Section 6 concludes our work and suggests future directions.

2 Related Work

Various forms of offensive language detection have recently attracted a lot of attention (Nobata et al., 2016; Pavlopoulos et al., 2017b; Park and Fung, 2017; Wulczyn et al., 2017). Apart from the growing volume of popular press concerning

[1] See, for example, https://goo.gl/VQNDNX.

[2] See also https://goo.gl/v7kA1K.

[3] https://www.perspectiveapi.com/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 571–576
Minneapolis, Minnesota, USA, June 6–7, 2019.

toxicity online, the increased interest in research into offensive language is partly due to the recent Workshops on Abusive Language Online,[4] as well as other fora, such as GermEval for German texts,[5] or TA-COS[6] and TRAC (Kumar et al., 2018),[7]. The literature contains many terms for different kinds of offensive language: toxic, abusive, hateful, attacking, etc. Largely, these are defined by different survey methods. In (Waseem et al., 2017), abusive language is divided into explicit vs. implicit, and directed vs. generalized. However, other researchers have created different taxonomies based on sub-kinds of toxic language (Table 2).

Although some previous research has considered several types of abuse and their relations (Malmasi and Zampieri, 2018), detecting varieties of hate has attracted more attention (Djuric et al., 2015; Malmasi and Zampieri, 2017; ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018). The first publicly available dataset for hate speech detection was that of Waseem and Hovy (2016). It contained 1607 English tweets annotated for sexism and racism. A larger dataset was published by Davidson et al. (2017), containing approx. 25K tweets collected by using a hate lexicon. Despite the popularity of hate speech detection in literature, no larger publicly available hate speech datasets seem to exist. For recent overviews of hate speech detection, consult Schmidt and Wiegand (2017) and Fortuna and Nunes (2018).

Research into the various kinds of offensive language detection is mainly focused on English, but some work in other languages also exists. Work on a large dataset of Greek moderated news portal comments is presented by Pavlopoulos et al. (2017a). A dataset of obscene and offensive user comments and words in Arabic social media was presented by Mubarak et al. (2017). Previous work includes a system to detect and rephrase profanity in Chinese (Su et al., 2017), and an annotation schema for unacceptable social media content in Slovene (Fišer et al., 2017).

[4] https://goo.gl/9HmSzc
[5] https://goo.gl/uZEerk
[6] http://ta-cos.org/
[7] https://goo.gl/DTZquU

3 Data

The SEMEVAL-2019 OFFENSEVAL dataset that is available to participants contains 13240 tweets; the counts of the labels are shown in Table 1. The OFFENSEVAL task consists of three subtasks, described in detail by Zampieri et al. (2019b). Subtask A aims at the detection of offensive language (OFF or NOT in Table 3). Subtask B aims at categorizing offensive language as targeting a specific entity (TIN) or not (UNT). Subtask C aims to identify whether the target of an offensive post is an individual (IND), a group (GRP), or unknown (OTH). Table 1 also shows the size of the vocabulary per class (label), which, unsurprisingly, is proportional to the class size. It is worth noting that offensive tweets targeting a group are the lengthier texts, with 28 tokens on average (see Table 1, column C, GRP column).

4 Baselines

We now describe the two baselines (**Perspective**, **BERT**) that we implemented and evaluated.

4.1 Perspective

We employed the **Perspective** API, which was created by Jigsaw and Google's Counter Abuse Technology team in Conversation-AI,[8] to facilitate better conversations online and protect voices in conversations (Hosseini et al., 2017). Although open-source code is available,[9] we chose to use pre-trained models, accessible through the API. For offensive language detection in Subtask A, we used the *Toxicity* model, which is a CNN based on GLOVE word embeddings,[10] trained over millions of user comments from publishers such as the New York Times and Wikipedia. This is a robust model, which we expect to be somewhat adaptable to different datasets (and their labels for closely related forms of offensive language), such as the offensive tweets of OFFENSEVAL. For offensive language categorization in Subtask B, we employed other experimental models, also available via the **Perspective** API, which detect various abuse types including those of Table 2.

4.2 BERT

BERT (Devlin et al., 2018) is a deep bidirectional network built using Transformers (Vaswani et al.,

[8] https://conversationai.github.io/
[9] https://goo.gl/yN196H
[10] https://goo.gl/rHYMqt

Subtask	A		B		C		
Label	NOT	OFF	UNT	TIN	IND	GRP	OTH
Number of Tweets	8840	4400	524	3876	2407	1074	395
Class specific vocabulary size	29.2K	18.6K	3.5K	17.3K	11.5K	7.8K	3.5K
Average number of tokens / Tweet	22	24	19	24	22	28	25

Table 1: Number of tweets, size of vocabulary, and average number of tokens per tweet, for each label (class). In Subtask A, the labels are 'not offensive' (NOT) or 'offensive' (OFF). In Subtask B, the labels are 'not targeted threat' (UNT) and 'targeted insult or threat' (TIN). In Subtask C, they are 'targeted insult or threat towards an individual' (IND), 'towards a group' (GRP), or 'towards another target' (OTH).

TOXICITY	@user Fuck you, you fat piece of shit
INSULT	Hey @user , you are disgusting.
THREAT	@user Kill the traitors.
PROFANITY	My wrist been fucked up for nearly a month now . This time im really going to the hospital to see what the fuck is wrong with it
IDENTITY ATTACK	Okay everyone always talks aboht the pathetic army and all the soy boy branches and gay shit and what not [...]
ATTACK ON COMMENTER	@user You are all utterly delusional. If you were really pro-life" you would [...]

Table 2: The tweets with the highest **Perspective** score per abusiveness type, on a trial dataset of 320 tweets shared by the competition organizers.

2017). It is pre-trained to detect (a) a masked word from its left and right context, and (b) the next sentence. We used the publicly available **BERT**-BASE version,[11] with 12 Transformer layers, 768 hidden states size, which is pre-trained on a monolingual corpus of 3.3B words. For a particular NLP task, a task-specific layer is added on top of **BERT**. In our case, the extra layer comprises dropout, a linear transformation, and softmax.[12] During the task-specific 'fine-tuning', the extra layer is trained jointly with **BERT** (refining the pre-trained **BERT** model) on task-specific data. Previous research demonstrated that fine-tuning **BERT** leads to state of the art performance in several NLP tasks (Devlin et al., 2018).

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
Perspective	**0.7933**	**0.8360**
BERT	0.7705	0.8163
BEST 2019	0.8290	—

Table 3: Results for Subtask A.

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	**0.8875**
All UNT baseline	0.1011	0.1125
Perspective	0.4785	0.6292
BERT	**0.6817**	0.8708
BEST 2019	0.7550	—

Table 4: Results for Subtask B.

5 Results

5.1 Offensive Language Detection

For Subtask A, we used the *toxicity* score from **Perspective** and returned the offensive label (OFF) when the returned score was above 0.5. No fine tuning was performed for **Perspective**. For **BERT**, we split the dataset to training (10K tweets) and development (3240) subsets, and fine-tuned **BERT** for 3 epochs.[13]

In this subtask, **Perspective** outperformed **BERT** and was ranked 12th out of 103 submissions. The difference from the top-ranked model was 3.5 F1 points. The performance of **Perspective** in this subtask is particularly interesting, considering that the training data for these models were not labeled for offensiveness, but rather for other attributes such as toxicity, threats, and insults.[14] Ignoring **Perspective**, **BERT** was ranked

[11] https://goo.gl/95mqhE

[12] We used default values for all hyper parameters.

[13] We used the uncased system with batch size 32, based on preliminary experiments.

[14] https://goo.gl/Bmiogb

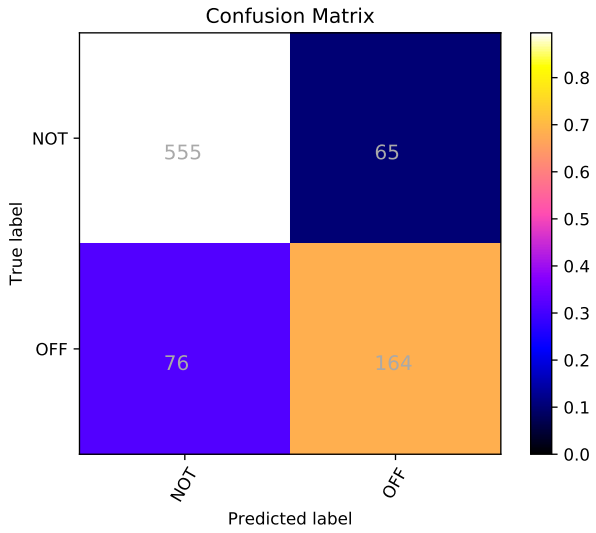

Figure 1: Confusion matrix for **Perspective** in Subtask A (Offensive Language Detection).

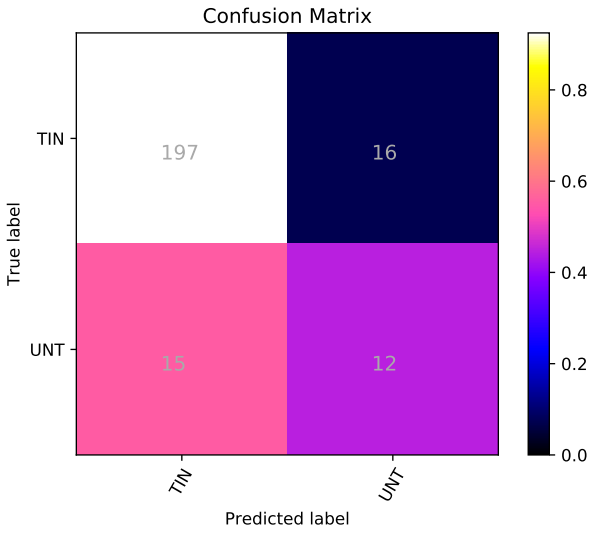

Figure 2: Confusion matrix for **BERT** in Subtask B (Offense Type Detection).

27th. As shown in Table 3, both of our strong baselines outperform the naive majority baselines for this subtask.

The confusion matrix of **Perspective** is shown in Fig. 1. Both recall and precision are high for the NOT label (87.96% and 89.81%), but lower for OFF (68.33% and 71.62%). This is explained by the fact that NOT is two times the size of OFF (Table 1). We also used **Perspective** to score the training data, since no fine-tuning was performed on the training data for **Perspective**. Macro F1 was 78.01% (85.02% for NOT, 71% for OFF) and accuracy was 80.24%, which are lower but close to the respective values on the test data (Table 3).

5.2 Offense Type Detection

For Subtask B, we used the experimental *insult*, *threat* and *attack on commenter* models from Perspective. We averaged *insult* and *attack on commenter* and used this average to compare with the *threat* score. The **Perspective** baseline returned a targeted insult/threat (TIN) when the average was greater, and untargeted (UNT) otherwise. The **BERT** baseline was fine-tuned on the entire dataset that was available to participants, because we considered that dataset too small for a training/development split.[15] **BERT** clearly outperformed the **Perspective** baseline (Table 4) and ranked 11th in this subtask among 73 participants, whereas the best system achieved 7.8 points higher in F1. The confusion matrix of **BERT** for this subtask is shown in Fig. 2. The large class imbalance (TIN tweets are approx. 7 times than UNT, see Table 1) significantly reduces both the recall (44.44%) and precision (42.86%) of **BERT** for the UNT class, compared to TIN (92.49% and 92.92%, respectively).

5.3 Offense Target Detection

For Subtask C, Perspective has no suitable model to respond yet and the **BERT**-based systems submitted were in an experimental phase, due to time constraints.[16] We consider the results we obtained for this subtask as not relevant and leave the development and evaluation of baselines for this subtask as future work.

6 Conclusion

This paper proposed and evaluated two strong baselines, based on the **Perspective** API and **BERT**, for identifying and categorizing offensive language in social media. Both baselines require few (**BERT**) or no additional task-specific training data (**Perspective**) and this is the first work, to our knowledge, to assess their performance in the tasks we considered. The **Perspective**-based baseline was ranked 12th among 103 submissions for the task of classifying a post as offensive or not. The **BERT** baseline was ranked 11th among

[15] We used the cased system with batch size 16, based on preliminary experiments.

[16] **BERT** base and **BERT** large (trained on CPU) were examined for this subtask, but preliminary experiments showed that the majority class was always returned.

73 submissions for the task of recognizing whether an offensive post is targeted or not. Both baselines were ranked surprisingly high in the corresponding tasks, considering that they were given no or few, respectively, additional task-specific training instances. Furthermore, the Perspective baseline, which required no fine tuning outperformed **BERT** by a large margin in the task of detecting offensive language. In future work, we intend to examine stronger, yet easy to apply baselines, and release source to make it easier to use them.

References

P. Burnap and M. L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515, Montreal, Canada.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

K. Dinakar, R. Reichart, and H. Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, pages 11–17.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *ICWWW*, pages 29–30.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint*.

D. Fišer, T. Erjavec, and N. Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable on-line discourse practices in slovene. In *1st Workshop on Abusive Language Online*, Vancouver, Canada.

P. Fortuna and S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

B. Gambäck and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *1st Workshop on Abusive Language Online*, pages 85–90, Vancouver, Canada.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. In *arXiv preprint*.

R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri. 2018. Benchmarking aggression identification in social media. In *TRAC*, Santa Fe, USA.

I. Kwok and Y. Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*, pages 1621–1622, Whasington, USA.

S. Malmasi and M. Zampieri. 2017. Detecting hate speech in social media. In *RANLP*, pages 467–472.

S. Malmasi and M. Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

H. Mubarak, K. Darwish, and W. Magdy. 2017. Abusive language detection on arabic social media. In *1st Abusive Language Workshop*, pages 52–56, Vancouver, Canada.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *ICWWW*, pages 145–153.

J. H. Park and P. Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *1st Workshop on Abusive Language Online*, pages 41–45.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2017a. Deep learning for user comment moderation. In *1st Workshop on Abusive Language Online*, pages 25–35.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *EMNLP*, pages 1125–1135, Copenghagen, Denmark.

A. Schmidt and M. Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain.

H.-P. Su, C.-J. Huang, H-T. Chang, and C.-J. Lin. 2017. Rephrasing profanity in chinese text. In *1st Workshop on Abusive Language Online*, Vancouver, Canada.

S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. In *TA-COS*, Portoroz, Slovenia.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *1st Workshop on Abusive Langauge Online*, Vancouver, Canada.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL SRW*, pages 88–93, San Diego, California.

E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex machina: Personal attacks seen at scale. In *ICWWW*, pages 1391–1399.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *NAACL*.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *SemEval*.

Z. Zhang, D. Robinson, and J. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *Lecture Notes in Computer Science*. Springer Verlag.