DeepAnalyzer at SemEval-2019 Task 6: A deep learning-based ensemble method for identifying offensive tweets

Gretel Liz De la Peña Sarracén and **Paolo Rosso**
PRHLT Research Center, Universitat Politècnica de València, Spain

Abstract

This paper describes the system we developed for SemEval 2019 on Identifying and Categorizing Offensive Language in Social Media (OffensEval - Task 6). The task focuses on offensive language in tweets. It is organized into three sub-tasks for offensive language identification; automatic categorization of offense types and offense target identification. The approach for the first subtask is a deep learning-based ensemble method which uses a Bidirectional LSTM Recurrent Neural Network and a Convolutional Neural Network. Additionally we use the information from part-of-speech tagging of tweets for target identification and combine previous results for categorization of offense types.

1 Introduction

The use of Internet has become an important media of personal and commercial communication. In this scenario, some users take advantage of the anonymity of this kind of communication, using this to engage in behaviour that many of them would not consider in real life. Therefore, much of the offensive language is widespread in social networks. Then, studying offensive language in texts from the social media is an essential task for security, the prevention of cyber-bullying, among other abusive behavior.

To increase the research in this areas, several workshops have been organized, such as ALW[1] and TRAC[2]. Recently, OffensEval[3] (Zampieri et al., 2019b), which is a shared task at the SemEval-2019[4] workshop has been launched on the research community. The aim of OffensEval is to deal with offensive language detection in the English language focusing on messages from Twitter.

In OffensEval, the treatment of offensive content is divided into three subtasks taking the type and target of offenses into account:

- A: Offensive language identification.
- B: Automatic categorization of offense types.
- C: Offense target identification.

In this work, we present the methodology proposed to each of these sub-tasks, which includes an ensemble of a LSTM Recurrent Neural Network and a Convolutional Neural Network, and additionally linguistic features for the last two subtasks. The architecture of the system will be more detailed in the following sections.

The paper is organized as follows. Next section briefly describes other works in this area. Then, Section 3 describes the proposed metodology and the dataset. Results are discussed in Section 4. Finally, we draw our conclusions together with a summary of our findings in Section 5.

2 Related Work

Some approaches have been proposed to tackle the problem of offensive language detection. It is the case of recent works (Waseem et al., 2017; ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018) and surveys (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018). There are even studies on languages other than English such as (Su et al., 2017) on Chinese and (Fišer et al., 2017) on Slovene.

Many of the last approaches rely on neural network models. For instance, the work of (Ganesan et al., 2018) presents a Multi-Layer Feedforward Neural Networks. Moreover, (Park and Fung, 2017) proposes to implement three models based

[1] https://sites.google.com/site/abusivelanguageworkshop2017/
[2] https://sites.google.com/view/trac1/home
[3] https://competitions.codalab.org/competitions/20011
[4] http://alt.qcri.org/semeval2019/index.php?id=tasks

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 582–586
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

on Convolutional Neural Networks (CNN) to classify sexist and racist abusive language: Char-CNN, WordCNN, and HybridCNN. It work reports that can boost the performance of simpler models. Also, (Pitsilis et al., 2018) proposes a detection scheme that is an ensemble of Recurrent Neural Network classifiers. It incorporates various features associated with user-related information. They report that the scheme can successfully distinguish racism and sexism messages from normal text.

3 Methodology and Data

The corpus provided by the organizers consists of 14,100 tweets in English. The data collection methods used to compile the dataset used in OffensEval is described in Zampieri et al. (2019a).

The first step is the preprocessing of the tweets, where texts are cleaned. All emoticons, hashtag and urls are removed. Then, the texts are represented as vectors with word embedding vectors. We used the pre-trained word vectors of Glove (Pennington et al., 2014), trained on 2 billion words from Twitter.

The method proposed in this work is based on an architecture that sequentially obtains the output for each of the subtasks. In the first level we use a model whose input is the word embeddings of a tweet and the output is a vector (*r_vector*) that is taken as a compact representation of the input and is used in the following steps. For the model, two types of networks have been used. In a first approach a Recurrent Neural Network (RNN) is used, and as a second approximation a Convolutional Neuronal Network (CNN). These two models are described below.

3.1 Convolutional Neural Network

The model is a version of the convolutional neural networks presented in (Kim, 2014) for sentence-level classification tasks. Here, the input of the model is a matrix where each row corresponds to the embedding vector of each word in the tweet. Three different filters of sizes 3, 4 and 5 are applied in a 1D convolution step to capture information from 3-grams, 4-grams and 5-grams. The feature maps produced by the convolution layer are forwarded to a Maxpooling layer. We used 2x2 filters for this pooling function on a feature map to reduce it to the single most dominant features.

Finally, the r_vector is generated by the concatenation of the results for each of the filters.

3.2 Recurent Neural Network

In NLP problems, standard LSTM Recurrent Neural Networks receive sequentially (left to right order) at each time step a word embedding w_t and produces a hidden state h_t.

On the other hand, the bidirectional LSTM makes the same operations as standard LSTM but, processes the incoming text in a left-to-right and a right-to-left order in parallel. Thus, the output is a two hidden state at each time step $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. The proposed method uses a Bidirectional LSTM network which considers each new hidden state as the concatenation of these two $\hat{h_t} = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The idea of this Bi-LSTM is to capture long-range and backwards dependencies.

3.3 Sub-task A

For the first sub-task, which consists in the identification of offensive language in tweets, *r_vector* is used as input of a Fully Connected Neural Networks (FCNN) of two layers with activation function relu. The class (offensive or not) is obtained in a third layer of two units, that refer to the number of classes, with a softmax activation function.

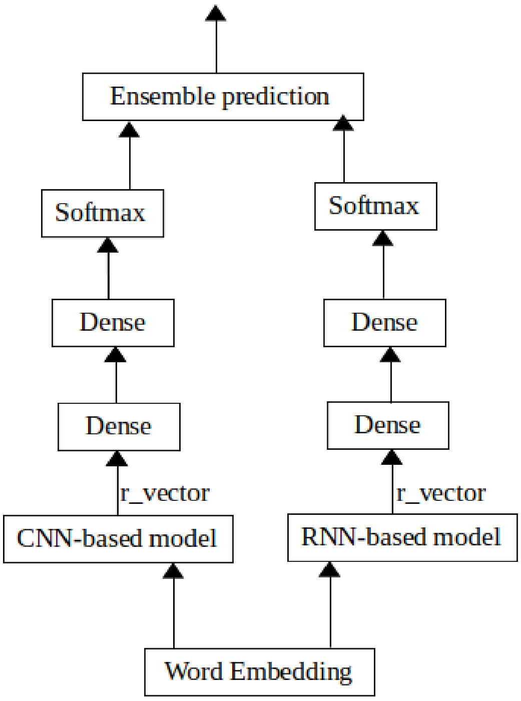

Figure 1: Architecture. Sub-task A

The Figure 1 shows the general scheme commented. Given this architecture, three weights of both CNN and RNN models were made. In the first weighting all the weight is for RNN (RNN_run). In contrast, in the second one, all the weight is for CNN (CNN_run). Finally, the third one is the actual ensemble model where both models are assigned equal weight (Ensemble_run). For combining the results of both models, the system gets the mean of the predictions of each one.

3.4 Sub-tasks B and C

In the sub-task of detecting the target of offensive language, the information of the part-of-speech tagging process of the tweets is used. This allows us to make more fine-grained distinctions on the words in texts which can identify to the target of aggressiveness. For instance, this information allows to discriminate between a proper noun and other kind of noun, and if a noun is plural or singular. In this way the model can learn sequences of tags which represent each type of target. The POS labels are obtained with Standford CoreNLP and they are represented as a one hot vectors. The sequence of labels is analyzed with a LSTM RNN, and a representation *p_vector* is obtained. Then, the concatenation of vectors *r_vector* and *p_vector* is used as input to another FCNN of one hidden layer with the activation function relu, and an output layer with two neurons with a softmax activation function. In this way, the prediction corresponding to the offensive target in the tweets is obtained. The Figure 2 shows this processing.

Finally, for the sub-task of classifying the types of offensive tweet, the prediction is obtained in a similar way to the previous sub-task. Here, a one hot vector corresponding to the POS tags present in the tweet is added to *r_vector*. Then, the prediction is calculated using another FCNN.

Finally, cross entropy is used as the loss function, which is defined as:

$$L = -\sum_i y_i * log(\hat{y}) \qquad (1)$$

Where y_i is the ground true classification of the tweet and $\hat{y}$ the predicted one.

4 Results

In the evaluation, the official ranking metric is macro-averaged F1. The results obtained in each subtask are shown in the next tables and confusion matrices. For each case, each of the three approaches discussed above (CNN_run, RNN_run and Ensemble_run) was evaluated and the results are shown in the tables with the name that was indicated. Also, random baseline generated by assigning the same labels for all instances are included. For example, ”All OFF” in sub-task A represents the performance of a system that labels everything as offensive. It was used for comparison.

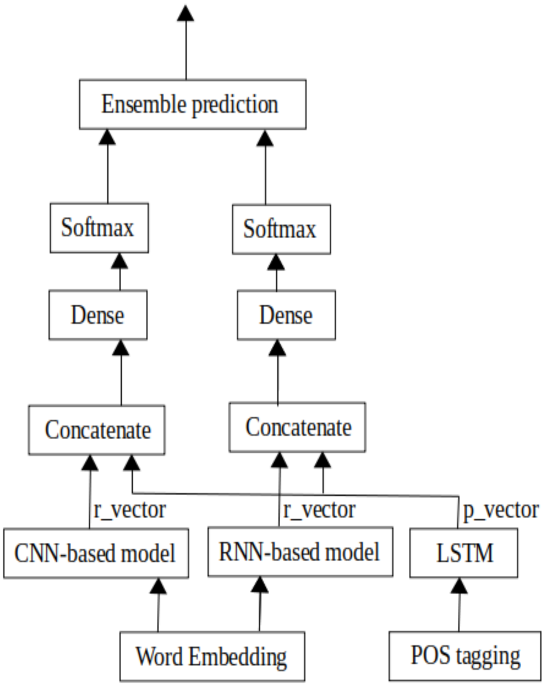

Figure 2: Architecture. Sub-task B

System	macro F1
All NOT baseline	0.4189
All OFF baseline	0.2182
Best	0.829
RNN_run	0.5984
CNN_run	**0.6600**
Ensemble_run	0.5925

Table 1: Results for Sub-task A

These results reveal a behavior not as good as expected, since although the baselines were exceeded in each case, the results were relatively far from the best results of the competition. Perhaps this is due to the fact that the different linguistic characteristics that could be extracted from tweets, such as information related to emoticons, hashtags and urls, were not analyzed in detail.

Another aspect to note is that for the three tasks

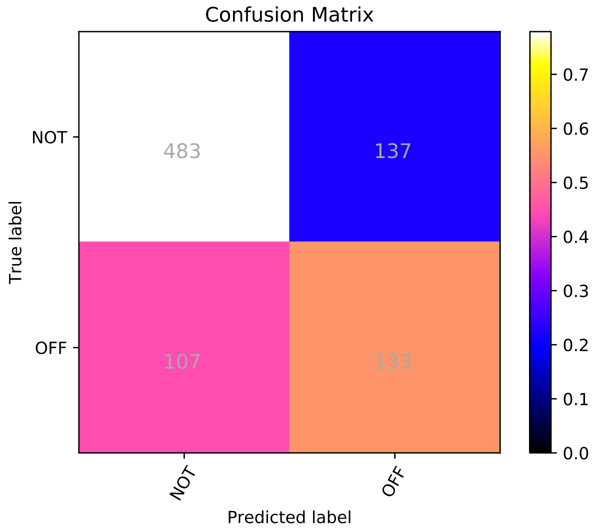

Figure 3: Sub-task A: CNN_run

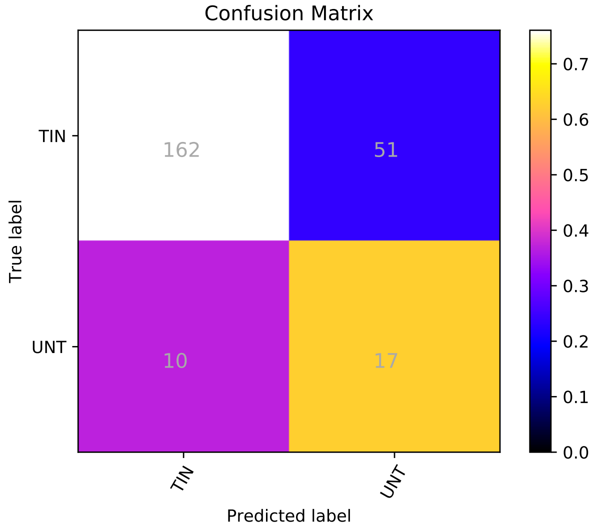

Figure 4: Sub-task B: RNN_run

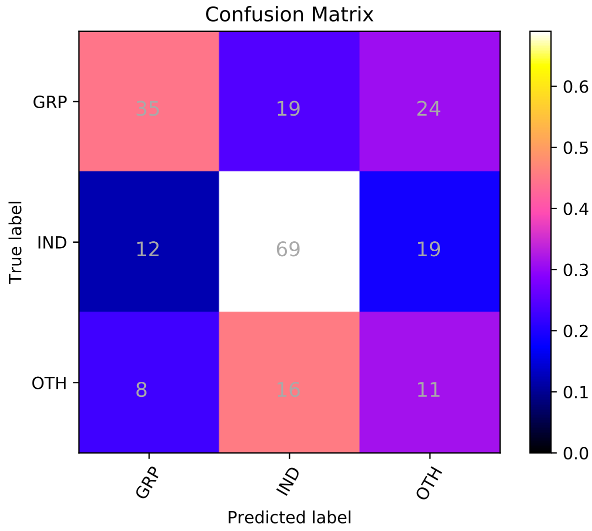

Figure 5: Sub-task C: CNN_run

System	macro F1
All TIN baseline	0.4702
All UNT baseline	0.1011
Best	0.755
RNN_run	**0.5997**
CNN_run	0.5704
Ensemble_run	0.5587

Table 2: Results for Sub-task B.

System	macro F1
All GRP baseline	0.1787
All IND baseline	0.2130
All OTH baseline	0.0941
Best	0.660
RNN_run	0.3848
CNN_run	**0.4833**
Ensemble_run	0.4174

Table 3: Results for Sub-task C.

the best approach is based on simple models instead of a combination of models that in our case was obtained with an ensemble of models based on neural networks. So that, for two of the tasks the best results were obtained only with the use of CNN and for the other one with the RNN.

5 Conclusion

In this paper our solution for the OffensEval challenge in SemEval 2019 was presented. We used an ensemble of models based on deep learning, and compared the results obtained to those ob- tained with each of the models independently. As a conclusion, it can be said that it may be more important for this kind of tasks to search for properly linguistic characteristics instead of designing complex models with a lot of parameters.

References

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Aishwarya Ganesan, Sneha Birajdar, Shivani Dalvi, and Jagruti Dandekar. 2018. Offensive language detection using ai technique.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

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.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

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 NAACL*.

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 (SemEval)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.