INF-HatEval at SemEval-2019 Task 5: Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter

Alison P. Ribeiro
Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brazil
alisonrib17@gmail.com

Nádia F. F. da Silva
Institute of Informatics
Federal University of Goiás
Goiânia – Goiás – Brazil
nadia@inf.ufg.br

Abstract

In this paper, we describe our approach to detect hate speech against women and immigrants on Twitter in a multilingual context, English and Spanish. This challenge was proposed by the SemEval-2019 Task 5, where participants should develop models for hate speech detection, a two-class classification where systems have to predict whether a tweet in English or in Spanish with a given target (women or immigrants) is hateful or not hateful (Task A), and whether the hate speech is directed at a specific person or a group of individuals (Task B). For this, we implemented a Convolutional Neural Networks (CNN) using pre-trained word embeddings (GloVe and FastText) with 300 dimensions. Our proposed model obtained in Task A 0.488 and 0.696 F1-score for English and Spanish, respectively. For Task B, the CNN obtained 0.297 and 0.430 EMR for English and Spanish, respectively.

1 Introduction

With the growth of users in social networks, there was also an increase in the odious activities that permeate these communicative structures. According to Nockleby et al. (2000), hate speech can be defined as any communication that deprecates a person or a group based on some characteristics such as race, color, ethnicity, gender, nationality, religion or other features. And the main motive that encourages users to spread hate on social networks is anonymity, so users can spread hate words to a particular target. For this reason, the hatred propagated can generate irreversible consequences, where young people who approach with cyberbullying and homophobia, mainly, commit suicide.

Nowadays, social networks like Twitter[1], Facebook[2] and YouTube[3] are pressured to develop tools to fight the proliferation of hate in their networks. A good example of this is the German government that threatened to fine social networks by up to 50 million euros if they did not fight the spread of hate (Gambäck and Sikdar, 2017).

However, while there is plenty of available content on social networks, the task of detecting hate speech remains difficult, largely because of the use of different sets of data for work, lack of benchmarking, and efficient approaches. Waseem, for example, bring a study focused on the detection of racism and sexism, whereas Nahar et al. 2012 and Sanchez and Kumar 2011) conducted a survey on detecting bullying. For the detection of homophobia, misogyny and xenophobia, the number of papers is still limited, one can cite a recent paper (Sanguinetti et al., 2018), where the authors sought to identify hate speech against immigrants. However, it is important that new research is publicized, because only in this way will it be possible to fight against hate in social networks.

Introducing a brief definition of hate speech and the importance of combating it, SemEval-2019 proposed a task in which it challenges participants to develop systems for detecting hate speech against women and immigrants on Twitter from a multilingual perspective , for English and Spanish.

The task was articulated around two related subtasks for each one of the languages involved: a basic task about hate speech, and another where refined hate content resources will be investigated to understand how existing approaches can handle the identification of especially dangerous forms of hatred, that is, those in which incitement is directed against an individual rather than against a group of people, and where aggressive behavior of

[1] https://twitter.com/

[2] https://www.facebook.com/

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

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 420–425
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

the perpetrator can be identified as a prominent feature of the expression of hatred. In order to reach this goal, this work proposed to develop a Convolutional Neural Network with the use of word embeddings.

The paper is organised as follows: previous work on hate-speech identification is discussed in Section 2. Section 3 presents details about the task, data sets and evaluation methods. Section 4 describes the methodology for categorizing hate speech based on deep learning, while experiments and results are reported in Section 5. Finally, Section 6 summarises the discussion.

2 Related Work

Some computational methods to detect hate speech are presented in this section. An example is the work of Badjatiya et al. (2017) that applied several algorithms of machine learning and deep learning, among them: Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosted Decision (GBDT), CNN and Long Short-Term Memory (LSTM). As a baseline they used char n-grams and bag-of-words, and as word embeddings they used GloVe and FastText. The objective was to classify if a tweet is racist, sexist or none, and the best result was 0.930 of F1-score, which was obtained through a LSTM model with Random Embedding and GBDT.

Another work that also follows a line of ternary classification was proposed by Malmasi and Zampieri (2017), where the purpose is to classify a tweet as hateful, offensive (but not hateful) or offensive language. For this, the researchers proposed an approach based on n-grams and word skip-grams using Support Vector Machine with cross-validation, the best result achieved was 0.78 of accuracy.

Gambäck and Sikdar (2017) developed a Convolutional Neural Network to classify hate speech on Twitter. In this case, the authors used 4 categories: racism, sexism, both (racism and sexism) and non-hate-speech. The structure of CNN was constructed with convolutional layers and pooling of 4 modes: character 4-grams, word vector based on semantic information built using word2vec (Mikolov et al., 2013a), randomly generated word vectors and word embeddings with character n-grams. In the classification phase, the softmax function and cross-validation with 10-folds were applied, the model based on word2vec embeddings best performed with 0.783 of F-score.

A recent study developed by Gaydhani et al. (2018) sought to address the difference between offensive language and hate speech, then the authors proposed several machine learning models based on n-grams and TF-IDF. The models were analyzed considering several n-values in n-grams and TF-IDF normalization methods. Consequently, the best result among several approaches was 0.956 of accuracy.

3 SemEval-2019 Task 5

In this section we will describe some details about data sets, tasks, and evaluation methods.

3.1 Dataset

The data for the task consists of 9000 tweets in English for training, 1000 for develop and 2805 for test. For Spanish, 4469 tweets for training, 500 for develop and 415 for test. The data were structured in 5 columns: id, text, Hate Speech (HS), Target Range (TR) and Aggressiveness (AG). See an example in the Table 1 (Basile et al., 2019). If the $@username$ is a woman, we have a case of feminicide.

id	text	HS	TR	AG
93874	@username stupid wish you die.	1	1	1
18267	Leftwing filth Deport them all. #Sendthemback	1	0	1
18345	1,500 migrants have died in Mediterranean in 2018	0	0	0

Table 1: Example of hate speech. Some examples are also taken from the data.

3.2 Task A

The task A is a two-class classification problem in which participants have to predict whether a tweet, in English or Spanish, with a particular target (women or immigrants) is hateful or not hateful – Hate Speech (1/0).

3.3 Task B

The purpose of this task is to: (i) classify hate tweets into English and Spanish, where tweets with hate speech, against women or immigrants, were identified as aggressive or non-aggressive, and (ii) identify the harassed target as just one person or group of individuals.

3.4 Evaluation

For the results evaluation of both tasks A and B, different metrics were used in order to allow more refined conclusions.

Task A. The systems will be evaluated according to the following metrics: accuracy, precision, recall and F1-score. The equations below show how the calculations are done. In the case of this task, the scores will be classified by F1-score. For better understanding, we will show the following definitions:

- *True positive* (TP): means a correct classification as odious. For example, the royal class is hateful and the model ranks as hateful.
- *True negative* (TN): means a correct classification as not hateful. For example, the royal class is not hateful and the model ranked as not hateful.
- *False positive* (FP): means a wrong classification as odious. For example, the royal class is not hateful and the model rated it as hateful.
- *False negative* (FN): means a wrong classification not hateful. For example, the royal class is hateful and the model ranked as not hateful.

$$\textit{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\textit{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Task B. In this task, the evaluation metrics are two: partial match and exact match. The strategy for the partial match is to evaluate the Hate Speech, Target Range and Agressiveness classes independently of each other using the metrics defined above. However, each system will include all measures and a summary of the performance in terms of macro-average F1-score, calculated according to the Equation 5. The exact match considers the predicted classes together, thus computing the Exact Match Ratio (Kazawa et al., 2005). Given the set of data consisting of n multi-label samples (Xi, Yi), where Xi denotes the i-th instance and Yi corresponds to the labels to be predicted (HS, TR and AG), the Exact Match Ratio (EMR) is calculated according to Equation 6.

$$\textit{F1-score} = \frac{F_1(HS) + F_1(AG) + F_1(TR)}{3} \tag{5}$$

$$\textit{EMR} = \frac{1}{n}\sum_{i=1}^{n} I(Yi, Zi) \tag{6}$$

where Zi denotes the set of labels predicted for the i-th instance and I is the indicator function.

4 Methodology

In this section, we describe the details of our proposed methods, including data preprocessing, neural networks and word embeddings.

4.1 Preprocessing

This step consists in eliminating noises and terms that have no semantic significance in classes prediction. For this, we performed the removal of links, numbers, special characters, and *stop words* (words with low discriminative power, for example, “is”, “that” etc.) and standardized in lowercase.

4.2 Word embeddings

Word Embeddings (Bengio et al., 2003) is a supervised statistical language model trained using deep neural networks. The purpose of this language model is to predict the next word, given the previous words of the sentence. The vector embeddings was a great advance in relation to the strategies based on the bag-of-words, which justifies its use in several works (Nakov et al., 2016; Poria et al., 2015; Cliche, 2017; Zhou et al., 2018; Rotim et al., 2017). For the proposed task, we use the GloVe (Pennington et al., 2014) and FastText (Joulin et al., 2016) model with 300 dimensions.

For the English language, we made use of the Stanford pre-trained GloVe (Pennington et al., 2014) where word embedding were trained with Wikipedia 2014 and Gigaword 5, while FastText (Joulin et al., 2016) was trained in Wikipedia 2017, UMBC webbase corpus and statmt.org news.

For the Spanish language, the GloVe vocabulary was computed from SBWC (Pennington et al., 2014; Cardellino, 2016), while FastText was computed from the Spanish Wikipedia (Bojanowski et al., 2016).

4.3 Convolutional Neural Networks

Initially, the Convolutional Neural Network architecture was designed for image processing, however it has been commonly used in the sentiment analysis (Wang et al., 2016; Cambria et al., 2016; Rosenthal et al., 2017; dos Santos and Gatti, 2014; Poria et al., 2015).

For the purpose of the task, a CNN was implemented based on the architecture proposed by (Zhang and Wallace, 2015), and this implementation can be divided in two steps: feature extraction and classification. In the feature extraction step, only two layers of convolution and two layers of pooling were used, with tanh activation function. Four filters were used, 2 of 3 dimensions and 2 of 4 dimensions. Each filter refers to the classic n-grams technique (extremely used in bag-of-words based models), which consists of processing a group of n words, in order to consider not only isolated words in a tweet, but also the context in which they are inserted. The filters are applied under the vector representation of the input tweets (embedding layer), using the concept of Back propagation to adjust the weights dynamically. According to Poria et al. (2016), these filters can extract lexical, syntactic or semantic features automatically. Finally, the two layers of convolution and pooling are concatenated and directed to the next step.

In the classification stage, we used two dense layers, the first one has 512 neurons, relu function of activation and dropout of 0.5. The second has a neuron and sigmoid activation function, phase where classification occurs. For the training, the loss and optimization functions used were binary_crossentropy and RMSprop (Hinton et al., 2012) (with learning rate 0.001), respectively.

5 Results

In this section we will discuss the results obtained by using a CNN for the detection of hate speech and the target of hate.

Task A				
English				
Model	**F1**	**P**	**R**	**Acc**
CNN-FastText	0.488	0.628	0.574	0.520
Spanish				
Model	**F1**	**P**	**R**	**Acc**
CNN-GloVe	0.696	0.708	0.712	0.696

Table 2: Results obtained related to Task A.

We obtained 0.488 of F1-score for English and 0.696 for Spanish with our CNN model using word embeddings, as shown in Table 2. This result also suggests that the combination of CNN and GloVe provides better results for this task.

English			
Class	**F1**	**P**	**R**
hateful	0.617	0.465	0.916
not hateful	0.359	0.792	0.232
Spanish			
Class	**F1**	**P**	**R**
hateful	0.685	0.598	0.802
not hateful	0.707	0.817	0.622

Table 3: Confusion matrix concerning task A.

The Table 3 displays the results of F1-score, Precision and Recall reached by class for each language. The F1-score can be used to measure the performance of the classifier, in this case CNN ranked the hateful class better, obtaining 0.617 of F1-score, while the result for not hateful class was 0.359 of F1-score in the English language.

From the perspective of the Spanish language, CNN obtained good results in the classification of both classes, hateful and not hateful, with 0.685 F1-score and 0.707 F1-score, respectively.

Task B		
English		
Model	**F1**	**EMR**
CNN-FastText	0.577	0.297
Spanish		
Model	**F1**	**EMR**
CNN-FastText	0.609	0.430

Table 4: Results obtained related to Task B.

Recapitulating the idea of task B, where the goal is to identify the target of the hate speech, that is, whether it is a single person or a group of individuals. Knowing that there is hate speech in the tweet (HS is 1), then one must detect if the target is only one person (TR is 1) or if it is a group of individuals (TR is 0), and if there is presence of aggressiveness in speech (AG is 1) or not (AG is 0). In this case, the EMR measure shows a percentage in which it corresponds to an accuracy rate, that is, it measures how much the model has managed not only to classify the hate speech, but also the target and the aggressiveness. The Table 4 shows the results of task B, where it was possible to obtain 0.297 EMR for English and 0.430 EMR for Spanish.

6 Conclusion

In this paper, we introduced the system that we proposed for SemEval-2019, task 5. Our goal was to experience an architecture that was adapted from a CNN using word embeddings. The task was to detect hate speech against women and immigrants on Twitter from a multilingual perspective, English and Spanish. We participate in two subtasks directed to the two languages, and we obtain the 18th position in the ranking of task A and the 19th position of task B in the English language. In the Spanish language, we obtain the 24th position in the ranking for both tasks.

The success of deep learning depends on finding an architecture to fit the task. Furthermore, as deep learning has scaled up to more challenging tasks, the architectures have become difficult to design by hand. In this paper, a CNN was implemented based on the architecture proposed by Zhang and Wallace 2015 and a fine-tuning of hyperparameters was not done for the proposed tasks (tasks A and B). In addition, other features were not exploited as sarcasm and irony, inherent in this type of domain. We intend to explore these and other features in future work.

Another discussion can be raised regarding the best performance to have happened in Spanish. The main hypothesis is related to the nature of the corpus used. It is observed that the test set of Spanish is smaller than that of English, besides being a corpus with "simpler texts to be classified" (Spanish texts have few signs of sarcasm). Such analyzes need further studies and will be evaluated in future work.

For future work as well, it would be interesting to explore systems that use different parameters for CNN and other word embeddings, such as Word2Vec (Mikolov et al., 2013b). It would also be interesting to construct an LSTM with attention mechanism proposed by Lin et al. (2017) and compare its performance.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, 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 (SemEval-2019)*. Association for Computational Linguistics", location = "Minneapolis, Minnesota.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings.

Mathieu Cliche. 2017. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*.

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.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda. 2005. Maximal margin labeling for multi-topic text categorization. In *Advances in neural information processing systems*, pages 649–656.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *CoRR*, abs/1712.06427.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Asia-Pacific Web Conference*, pages 767–774. Springer.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1–18.

John T. Nockleby, Kenneth L. Karst Leonard W. Levy, and Adam Winkler. 2000. *Hate Speech. In Encyclopedia of the American Constitution*. New York : Macmillan Reference USA, ©2000.

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.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Leon Rotim, Martin Tutek, and Jan Šnajder. 2017. Takelab at semeval-2017 task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 866–871.

Huascar Sanchez and Shreyas Kumar. 2011. Twitter bullying detection. *ser. NSDI*, 12:15–15.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of LREC*.

Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 225–230.

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.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Liyuan Zhou, Qiongkai Xu, Hanna Suominen, and Tom Gedeon. 2018. Epution at semeval-2018 task 2: Emoji prediction with user adaption. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 449–453.