Know-Center at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter using CNNs

Kevin Winter
Know-Center GmbH
Inffeldgasse 13
Graz, 8010, Austria
kwinter@know-center.at

Roman Kern
Know-Center GmbH
Inffeldgasse 13
Graz, 8010, Austria
rkern@know-center.at

Abstract

This paper presents the *Know-Center* system submitted for task 5 of the SemEval-2019 workshop. Given a Twitter message in either English or Spanish, the task is to first detect whether it contains hateful speech and second, to determine the target and level of aggression used. For this purpose our system utilizes word embeddings and a neural network architecture, consisting of both dilated and traditional convolution layers. We achieved average F1-scores of 0.57 and 0.74 for English and Spanish respectively.

1 Introduction

The ever-increasing number of message board forums, social media platforms and other websites that allow user comments enable participants to express their opinions freely and sometimes even anonymously. This barley restricted access in combination with the unmanageably vast amount of user-generated content unfortunately also creates an environment, which is vulnerable to profanity and hateful speech, rendering it hostile to the individuals or groups of people targeted. This problem is of increasing importance (Kettrey and Laster, 2014) and calls to establish systems that allow for automated detection of such behavior.

While detecting abusive language by it self is already a challenging task, (Malmasi and Zampieri, 2018) showed that it is even harder to differentiate between its subtypes. Messages containing profanity, sexism, racism and other forms of hateful speech may be formed using very similar vocabularies. Additionally, these subtypes may overlap, making data sets dependent on the subjective judgments of annotators. This may be particularly true for finding a threshold distinguishing aggressive and none-aggressive speech. In a task organized as part of COLING 2018 (Kumar et al., 2018) 15,000 participants were asked to detect aggressive behaviour in a data set of Facebook posts. Even the best system only obtained a weighted F1-score of 0.64.

Task 5 of the SemEval-2019 workshop (Basile et al., 2019) aims to accelerate the research and development of such systems. It comprises two distinct subtasks. The first subtask is devoted to detect hateful speech against immigrants and women in Twitter messages. For the second subtask, messages that are detected to be hateful are investigated further. Here, the goal is to identify aggressive behaviour and the target, women or immigrant, harassed. Given the multilingual nature of the task, systems for both English and Spanish are required. For the design and evaluation of these systems, training data sets for both languages are provided. The English data set consists of 9000 annotated messages whereas the Spanish data set consists of 4500.

Closely related, another task is hosted at the SemEval-2019 workshop, dealing with offensive language in social media and the individuals and groups targeted by it (Zampieri et al., 2019).

The rest of the paper is organized as follows. In the next section we give an overview of work that has been done in the field of detecting hateful and abusive speech. Section 3 describes our submitted system, including the pre-processing performed as well as the classifier trained on the data sets provided. In section 4 we show the results of our system and compare them with those of other participants in this challenge. Section 5 contains a brief summary.

2 Related Work

The problem of abusive behaviour in online media has been addressed in various fields. (Olteanu et al., 2018) address the causes for hateful speech

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 431–435
Minneapolis, Minnesota, USA, June 6–7, 2019.

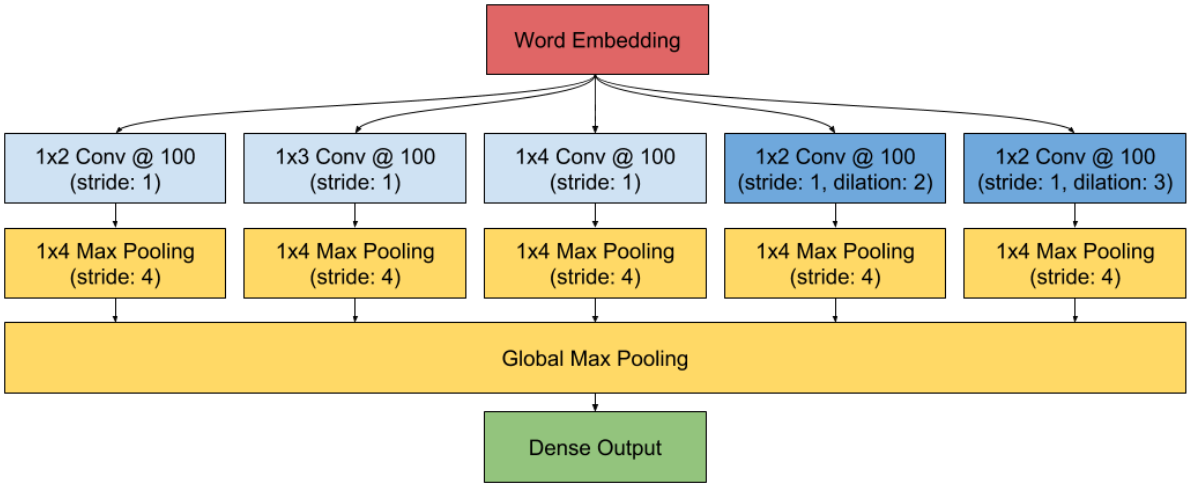

Figure 1: Network architecture of our classifier. Our system makes use of different strategies of applying convolutions, including dilated convolutions.

and in particular the effects of violent attacks performed by extremist groups and individuals. They were able to show, that such events majorly fuel hateful comments on Twitter and Reddit.

For the detection and classification task multiple methods have been previously explored. One approach involves the identification of words and parts of words in the form of character n-grams, that are the most indicative of hateful speech (Waseem and Hovy, 2016; Nobata et al., 2016). A potential downside of such approach might be that the meaning of a word may depend on the context it is used in (Sood et al., 2012). Chen et al. (2012) tried to overcome this issue by employing word n-grams. However, this is associated with an increase in feature space.

Another approach is to employ word embeddings in order to capture similarities between words (Badjatiya et al., 2017). Extending this method, Djuric et al. (2015) used paragraph2vec (Le and Mikolov, 2014) to encode whole messages and detect hateful messages in the embedding vector space. In addition to these lexical features, linguistic and syntactic features can be extracted. These may include the length of messages, average length of words, number of punctuations, Part-of-Speech (PoS) tags and dependency relationships (Nobata et al., 2016).

Recent work on this topic include different neural network architectures to classify text. Here either recurrent neural networks (RNNs) like LSTMs and GRUs, convolutional neural networks (CNNs) or both have been researched (Zhang and Luo, 2018; Zhang et al., 2018; Badjatiya et al., 2017).

3 System Description

Following the latest developments in the field of text classification, our system utilizes different convolutional filters in order to extract features. In the following sections we will describe the steps performed to pre-process the raw text messages and the network architecture used.

3.1 Pre-Processing

Given the raw Twitter messages, several steps of pre-processing are applied. First we follow the suggestions of Pennington et al. (2014), which involves the replacement of certain characters by tags. User mentions are replaced by `"<user>"`, numbers by `"<number>"`, web links by `"<url>"` and repeating characters like `"!!!"` by `"! <repeat>"`. Furthermore, words that are written using uppercase characters only are replaced by the same word in lowercase, followed by `"<allcaps>"`. Hashtags like `"#IllegalImmigrants"` are replaced with `"<hashtag> illegal immigrants"`, splitting the text before each uppercase character. The resulting text is then padded with `"<space>"` to match the length of the longest message.

The sequences are then encoded using pre-trained word embeddings. For English, the 200-dimensional GloVe embeddings from Pennington et al. (2014) are used, because they

were trained on Twitter messages and include vectors for the tags mentioned above. For Spanish, the 300-dimensional GloVe embeddings trained on the Spanish Billion Word Corpus (Cardellino, 2016) are used. Unknown words were replaced by "<unknown>" in English and "desconocido" in Spanish. These embeddings are then directly fed to our neural network classifier.

In order to compare our model to traditional approaches as discussed above, we also implemented a second pre-processing pipeline. This takes the padded sequences, applies stemming and extracts *n*-grams tuples with *n* being in the range from one to four words. These tuples are then vectorized using *TF-IDF*.

3.2 Classification

The detection of hate speech as well as the classification of aggressive behavior and the target harassed can be seen as three independent binary classification tasks. Hence, we can use the same model for each individual subtask and language. The only exception to this is the size of the input, since the the English embedding is 200-dimensional, whereas the Spanish embedding is 300-dimensional.

The network itself employs five different filter types, as shown in Figure 1. All of them are 1-d convolutions, meaning that the windows spread along all dimensions of the embedding size and only move along the words in a message. The first three are traditional convolutional filters with a window size of two, three and four words. These can be seen as n-gram feature extractors. The other two filters have a window size of two, but are dilated with dilation rates of two and three. By skipping words in between two other words, these filters may extract word combinations that may otherwise be missed due to the low importance of the words in the middle. All filter types use a stride of one, same padding and rectified linear units (ReLUs) as activation functions. For each of the five filter type, 100 filters are used. Max-pooling with a filter size of four and a stride of four is performed on all, to reduce dimensionality. In the next layer, the filter maps are concatenated and global max-pooling is performed. This flattens the filters in a non-parametric way and extracts the most pronounced features along all filters. Finally, these features are fully connected with one output neuron, which uses the sigmoid activation function. The network is trained for ten epochs using the Adam optimization algorithm (Kingma and Ba, 2014) and a batch size of 32.

Model	EN	ES
CNN + DIL	**0.78**	**0.82**
CNN	0.77	0.80
SVM	0.63	0.68
LogReg	0.65	0.68
SVM (*n*-gram *TF-IDF*)	0.76	0.81
LogReg (*n*-gram *TF-IDF*)	0.76	0.77

Table 1: Comparison of model performance (F1-score) on the training set for subtask 1 (hate speech detection).

We found this model to yield the best performance, comparing it to various other approaches. For the test set we selected ten percent of the training data points randomly and stratified. The results of this comparison can be seen in table 1. In order to evaluate the effectiveness of dilated convolutions we removed them from the model, leaving just the three regular convolutional filter types. As a result the F1-scores dropped by 0.01 in English and 0.02 in Spanish. Furthermore, logistic regression and SVM classifiers were trained both on the word embeddings and the *n*-gram based *TF-IDF* features. On the training data our approach performs slightly better than these. One reason for this very marginal improvement over the traditional approaches may be the size of the training set. In order to train a model with a high number of parameters like ours large data sets are majorly beneficial.

4 Results

Here we show the performance of the *Know-Center* system on the challenge's official test sets and compare it with the performances of the other participants. The results are shown in Tables 3 and 2 for the English and Spanish tasks respectively. The provided rankings refer to the average F1-scores over all subtasks, namely hate speech detection, target classification and aggression classification.

As illustrated, the *Know-Center* system achieved F1-scores of 0.45 and 0.72 in identifying hate speech in English and Spanish. For the target classification, we achieved F1-scores of 0.69 and 0.81. In detecting aggressive behavior, our system

#	Team name	AVG	HS	TR	AG
1	MITRE	0.77	0.76	0.82	0.73
2	Saagie	0.76	0.72	0.81	0.76
3	Atalaya	0.76	0.74	0.81	0.73
4	CIC-2	0.76	0.73	0.80	0.74
5	INGEOTEC	0.75	0.71	0.82	0.74
10	**Know-Center**	**0.74**	**0.72**	**0.81**	**0.70**
15	SVC baseline	0.74	0.70	0.78	0.73
26	MFC baseline	0.40	0.37	0.42	0.41

Table 2: F1-scores for each subtask in Spanish (subtask 1: hate speech detection (HS), subtask 2: target (TR) and aggression (AG) classification) as well as the overall average (AVG) per team, including their rank (#)

#	Team name	AVG	HS	TR	AG
1	scmhl5	0.63	0.60	0.71	0.59
2	alonzorz	0.61	0.52	0.75	0.57
3	MITRE	0.61	0.53	0.74	0.58
4	SINAI-DL	0.61	0.52	0.71	0.60
5	YNU_NLP	0.61	0.50	0.71	0.62
17	SVC baseline	0.58	0.45	0.70	0.59
20	**Know-Center**	**0.57**	**0.45**	**0.69**	**0.57**
42	MFC baseline	0.42	0.37	0.45	0.45

Table 3: F1-scores for each subtask in English (subtask 1: hate speech detection (HS), subtask 2: target (TR) and aggression (AG) classification) as well as the overall average (AVG) per team, including their rank (#)

achieved F1-scores of 0.57 and 0.70.

This is particularly interesting when comparing these results with the once obtained on the training data. Here the Spanish model performs similarly, but the English model does not. In general, the F1-scores obtained in the Spanish subtasks are better across all teams, even though less teams participated in it. One reason for that may be differences between training and test set. Besides that, our model uses high dimensional features, given the size of the word embeddings and the message length. For this, the size of the training set is very small, which makes it difficult to train a model with a high number of parameters such as ours. Even though a validation set was used during training, the possible homogeneity of the training set may have led to an over-fitting of the model.

5 Conclusion

We framed the tasks an a binary text classification problem, for which we developed a classification method that we used to participate in both subtasks of task 5 of the SemEval-2019 workshop. The classifier makes use of word embeddings and CNNs to identify hate speech in Twitter messages, determine the target and aggressive behavior. The same pre-processing and network architecture has been used for all tasks and languages. Averaged over all subtasks we achieved F1-scores of 0.57 and 0.74 for English and Spanish respectively.

Acknowledgments

The Know-Center GmbH Graz is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

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.

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

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Heather Hensman Kettrey and Whitney Nicole Laster. 2014. Staking territory in the world white web an exploration of the roles of overt and color-blind racism in maintaining racial boundaries on a popular web site. *Social Currents*, 1(3):257–274.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. 2012. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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).

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *arXiv preprint arXiv:1803.03662*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.