Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

InriaFBK Drawing Attention to Offensive Language at Germeval2019

Michele Corazza[†], **Stefano Menini**[‡]
Elena Cabrio[†], **Sara Tonelli**[‡], **Serena Villata**[†]
[†]Université Côte d'Azur, CNRS, Inria, I3S, France
[‡]Fondazione Bruno Kessler, Trento, Italy
michele.corazza@inria.fr
{menini,satonelli}@fbk.eu
{elena.cabrio,serena.villata}@unice.fr

Abstract

In this paper we describe the system developed by InriaFBK team and submitted to the Germeval2019 task on offensive language detection and classification. With the same architecture we participate to all subtasks: binary classification of offensive and not offensive tweets, 4-class message categorisation based on offense type (Profanity, Insult, Abuse and Other), and classification of explicit and implicit offensive language. The two runs submitted for each subtask are obtained with and without attention mechanism. After evaluating our system performance on Germeval2018 test set, we observe that attention is remarkably beneficial in the more challenging tasks of implicit offense detection and offense categorisation.

1 Introduction

Detecting hurtful, derogatory and obscene comments online has become of paramount importance for the well-being of users, who access social networks to exchange ideas and build a sense of community, as well as for social media platforms, which have been accused of fostering the widespread of hurtful content. Recent initiatives at institutional level have been undertaken to limit the phenomenon of online hate speech, see for example the 2018 Code of Conduct signed between EU representatives and four major social media players[1]. The monitoring process following the adoption of this code of conduct has shown that, when users report offensive content, 88.9% of them get a reply from the social media platform within 24 hours. However, if we compare Facebook, YouTube, Instagram and Twitter, statistics show that the latter tends to remove remarkable less content than the others, i.e. only 43.5% of reported messages compared to 71.7% by the other three platforms on average.[2] The EU report highlights how most of the feedback from Twitter is on *trusted reports* rather than on general users reports, making reasonable to think that Twitter policies are less restrictive and do not aim to comply with every user. This makes the issue of automatic hate speech detection on Twitter even more urgent, especially when developed systems are able to identify different types of offense and cope with implicit hate messages.

In this paper, we present our system submitted to the Germeval 2019 task for offensive language detection, and detail the two runs for each of the three subtasks (with and without attention mechanism). The general framework is an improved version of the InriaFBK system developed for Italian hate speech detection (Corazza et al., 2018a) and for German at Germeval 2018 (Corazza et al., 2018b), with a more careful choice of external embeddings and of neural network parameters. Furthermore, we evaluate the contribution of attention mechanism to each of the subtasks.

2 Related work

Many solutions and resources are available to perform hate speech detection and classification on English data. For example, in Waseem and Hovy (2016) the authors not only present their work on the classification of racist and sexist tweets adopting a logistic regression model based on one-to-four-character n-grams, but also release an annotated dataset for the task. Other classification approaches have been tested on the same dataset, see for example (Kshirsagar et al., 2018) proposing a neural classifier using pre-trained word embeddings and max/mean pooling from fully-

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32008F0913&from=EN

[2] https://ec.europa.eu/commission/news/countering-illegal-hate-speech-online-2019-feb-04_en

connected transformations of these embeddings. The approach is tested also on the the HATE dataset (Davidson et al., 2017) and on the Harassing dataset (Golbeck et al., 2017) showing that it is able to learn associations among words typically used in hateful communication.

Other approaches targeting languages different from English have been proposed mainly in the context of shared tasks, such as the first *Hate Speech Detection* (HaSpeeDe) task for Italian (Bosco et al., 2018) and *Aggressiveness Detection* (Carmona et al., 2018) for Mexican Spanish at IberEval 2018.

As for German, the first shared task was proposed in 2018 on the *Identification of Offensive Language* (Wiegand et al., 2018). The task covers the detection of offensive comments from a set of German tweets, that had to be further classified into abusive language, insults and profane statements. The systems presented by the participants introduce a number of different approaches, ranging from feature-based supervised learning (i.e., SVMs for the top-performing system TUWienKBS (Padilla Montani and Schüller, 2018)) to deep learning. Most top performing systems in both subtasks are based on deep learning, such as spMMMP (von Grunigen et al., 2018), uhhLT (Wiedeman et al., 2018), SaarOffDe (Stammbach et al., 2018), InriaFBK (Corazza et al., 2018b).

Looking at the systems participating in the above tasks, we observe a number of features shared by many deep learning approaches, such as domain-specific word embeddings, the use of emotion or sentiment lexica, features related to the message (e.g. length, punctuation marks, etc.) as well as specific pre-processing steps.

3 Data and Tasks

At the Germeval evaluation, three different subtasks were proposed: one for the detection of offensive messages, one for a fine-grained classification in four classes, namely *Profanity*, *Insult*, *Abuse* and *Other*, and one for the identification of explicit and implicit hate. Since subtask I and II had already been proposed at Germeval2018, the organisers allowed participants to use as training data the data released both in 2018 and 2019 as training data. For the third subtask, instead, the dataset was novel.

For the submissions of subtask I and II, we use the concatenated Germeval 2018 and 2019 training sets for training and the Germeval2018 test data as validation set. For subtask III we isolate 20% of the training set for validation (see details in Section 4). Below we summarise the number of instances used as training for each subtask:

Subtask I - Binary classification: The two labels are 'offensive' and 'other'. The latter was reserved for tweets which were not offensive. The binary classification subtask involved 2,975 messages with 'offensive' label and 6,029 messages with the 'other' label.

Subtask II - Fine-grained classification: The four classes annotated are 'profanity', 'insult', 'abuse' and 'other'. In the corpus, there are 1,220 messages for 'insult', 223 for 'profanity', 1,532 for 'abuse', and 6,029 messages for 'other'.

Subtask III - Implicit and Explicit offense classification: All messages in this dataset are offensive, but they are labeled either as 'implicit' or 'explicit'. In particular, there are 1,699 explicitly offensive message, and 259 implicit ones.

To compute the preliminary evaluation results reported in this paper, instead, we change the splits by using 20% of the 2018 and 2019 training sets for subtasks I and II for validation, and compute the performance reported in the following tables on the Germeval 2018 test set. For subtask III, we use 20% of the training set for validation and 20% as test set.

4 System Description

In order to perform an analysis of the activations of an attention mechanism, we use a recurrent neural network with attention applied to the outputs of the recurrent GRU layer, and compare its performance to the same network with no attention applied. Since the domain of the task is interactions on a social media platform, we apply some ad-hoc preprocessing steps, which are detailed in the next subsection, in order to improve the performance of the classifier on Twitter-specific language.

To isolate the validation set from the training data, we use `train_test_split` from scikit-learn (Pedregosa et al., 2011).

4.1 Preprocessing

Since the language of social media interactions presents unique challenges for standard NLP tasks, we normalise the tweet content by replacing user mentions and URLs with the strings "username"

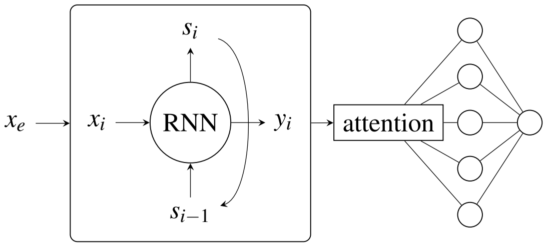

Figure 1: The recurrent neural architecture

and "URL" respectively. We do not apply hashtag splitting, since it proved not effective on German in a comparative evaluation for hate speech detection (Corazza et al., 2019).

4.2 Word embeddings

Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are a widely used approach to represent word meaning in natural language processing tasks, as they allow to acquire some information about words through an unsupervised process. However, word embedding resources have a major drawback when it comes to processing German data, since they may not contain all compounds or all the declinations of a single word, resulting in many out-of-vocabulary terms. This issue can be alleviated by using subword information to represent a term as the sum of the vectors representing its character n-grams. This is the main reason why we chose to use FastText embeddings (Bojanowski et al., 2016), pretrained on Common Crawl and Wikipedia [3].

4.3 Recurrent model

We develop a simple recurrent neural network model and use it for all subtasks. We use the word embeddings from the words of each tweet as input for a GRU (Cho et al., 2014) of size 100. Recurrent dropout of 0.2 is applied to the GRU. The output at the last timestep from the GRU is then fed to a single, fully-connected layer with 200 neurons, followed by one or more output neurons, depending on the subtask. For subtasks 1 and 3 a single, sigmoid activated neuron is used, while for subask 2 we use four outputs with a softmax activation. The binary subtasks use binary crossentropy as the loss function, while subtask 2 uses categorical crossentropy. The optimizer used is Adam and the models were implemented in Keras (Chollet and others, 2015). In addition to classifying offensive language, our goal was also to examine how an attention mechanism could improve performance, and whether the activations could be used to understand the classifier behavior. In particular, we consider the output of the GRU layer g:

Category	Precision	Recall	F1 Score
No Attention			
Offensive	0.677	0.630	**0.653**
Other	0.831	0.859	0.845
Macro AVG	0.754	0.744	**0.749**
+ Attention			
Offensive	0.692	0.595	0.640
Other	0.821	0.875	**0.847**
Macro AVG	0.757	0.735	0.746

Table 1: Subtask 1 without attention (above) and with attention (below)

$$GRU(x) = (e_1, e_2, \ldots, e_n) \quad e_i \in \mathbb{R}^{100} \qquad (1)$$

We then apply a perceptron layer to each of the outputs of the GRU and use softmax to obtain weights that sum to 1:

$$A(e) = softmax(F(e)) \qquad (2)$$

Where:

$$\begin{gathered} F(e) = (f(e_1), \ldots, f(e_n)) \\ f(e_i) = (We_i + b) \\ W \in \mathbb{R}^{1\times 100} \quad b \in \mathbb{R}^1 \end{gathered} \qquad (3)$$

After applying a perceptron layer to each output of the GRU, we use a softmax layer so that the sum of all timesteps is one (padding is ignored). The weights obtained are then multiplied elementwise with the outputs of the GRU:

$$a(x) = A(GRU(x)) \odot GRU(x) \qquad (4)$$

We then sum over the vectors obtained by applying attention to the outputs of the GRU, and use the resulting vector to classify offensive language, by feeding it to a single, fully connected hidden layer followed by the outputs.

5 Evaluation

For subtasks I and II, we report below the results obtained on the Germeval 2019 test set, comparing the system performance with and without attention mechanism.

With respect to subtask I (see Table 1), the two models perform similarly well. In particular, while the model without attention is the better performing one with respect to the offensive class, the F1

[3] https://fasttext.cc/docs/en/crawl-vectors.html

Category	Precision	Recall	F1 Score
No Attention			
Abuse	0.371	0.455	0.409
Insult	0.375	0.397	**0.385**
Profanity	0.355	0.099	0.155
Other	0.832	0.817	0.825
Macro AVG	0.483	0.442	0.462
+ Attention			
Abuse	0.443	0.503	**0.470**
Insult	0.425	0.325	0.367
Profanity	0.475	0.171	**0.252**
Other	0.824	0.874	**0.849**
Macro AVG	0.542	0.468	**0.502**

Table 2: Subtask II without attention(above) and with attention mechanism (below)

metric on the ”other” class is remarkably similar, with a slight advantage for the model with attention. This results in a slight advantage in terms of macro average F1 for the model without attention.

For Subtask II (see Table 2), focusing on fine-grained classification, the observed behaviour of the two models is still similar, but this time the attention-based model outperforms the attention-less one across all categories except for the ”insult” one, showing that attending over single words can be useful when classifying different types of offensive language. The largest improvement is achieved on the Profanity class, showing that attention mechanism in this case can better learn from few examples (only 223 for this class), while it is less evident on the Other class, which is the majority one (6,029 training instances).

Category	Precision	Recall	F1 Score
No Attention			
Explicit	0.891	0.964	**0.926**
Implicit	0.580	0.299	0.394
Macro AVG	0.735	0.631	0.679
+ Attention			
Explicit	0.910	0.918	0.914
Implicit	0.488	0.463	**0.475**
Macro AVG	0.699	0.690	**0.695**

Table 3: Subtask III without attention (above) and with attention mechanism (below)

With respect to subtask III (see Table 3), we observe a more significant difference between the two models on the Implicit class, while the Explicit one is equivalent. This may confirm the model behavior observed in subtask II, where classes with less examples had improved performance when using attention. Also in this case, the model using attention has a higher F1 score value for the implicit class, for which only 259 training instances are available.

6 Attention activations

In order to understand how attention affects the classification outcome and whether the outputs of the attention layer can help explain the classification performed by our model, we examined the attention for each word in the test set of Germeval 2018 (the outputs of Equation 2), using a model trained on the first subtask. Looking at the lemmas ranked by average weights learned by the attention mechanism, we observe that the top ones are mostly emotionally loaded with a negative connotation. For example, we find among the lemmas with highest weights words such as *klatsch*, *Opportunistin*, *Elektrojude*, and *verpisst*. Looking at attention weights of the words composing a tweet, we observe the same trend: in the messages correctly classified as ‘Offensive’ the words with highest attention weights are those with negative polarity, that mostly contribute to correct classification. For example, in *Von mir aus könnt ihr jämmerlich verrecken* the last two words have the highest attention. In a similar way, the last word of the following tweet is the one with highest attention weight: *Ich persönlich scheisse auf die grüne Kinderfickerpartei*. These findings suggest that the attention mechanism is effectively capturing the words whose meaning and polarity most contribute to the classifier choice. Furthermore, examining activation weights can lead to precious insight into the inner criteria used by models to detect offensive language.

7 Conclusions

In this work we detailed the system runs submitted by the InriaFBK team to Germeval 2019. With the same architecture we participated in all three subtasks, performing both binary and multi-class classification. In a comparative evaluation, our results show that the attention mechanism has a positive impact on classes with few training instances, while it has no remarkable effect on classes that are well represented in the training set.

Acknowledgments

Part of this work was funded by the CREEP project (`http://creep-project.eu/`), a Digital Wellbeing Activity supported by EIT Digital in 2018. This research was also supported by the HATEMETER project (`http://hatemeter.eu/`) within the EU Rights, Equality and Citizenship Programme 2014-2020.

References

[Bojanowski et al.2016] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

[Bosco et al.2018] Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

[Carmona et al.2018] Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 74–96.

[Cho et al.2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

[Chollet and others2015] François Chollet et al. 2015. Keras. `https://keras.io`.

[Corazza et al.2018a] Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018a. Comparing different supervised approaches to hate speech detection. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

[Corazza et al.2018b] Michele Corazza, Stefano Menini, Arslan Pinar, Rachele Sprugnoli, Cabrio Elena, Sara Tonelli, and Villata Serena. 2018b. Inriafbk at germeval 2018: Identifying offensive tweets using recurrent neural networks. In *Proceedings of GermEval 2018*, pages 80–84.

[Corazza et al.2019] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Robust hate speech detection: A cross-language evaluation. *Under review.*

[Davidson et al.2017] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 512–515.

[Golbeck et al.2017] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 229–233.

[Kshirsagar et al.2018] Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32. Association for Computational Linguistics.

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

[Padilla Montani and Schüller2018] Joaquin Padilla Montani and Peter Schüller. 2018. Tuwienkbs at germeval 2018: German abusive tweet detection. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, 09.

[Pedregosa et al.2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pennington et al.2014] 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.

[Stammbach et al.2018] Dominik Stammbach, Azin Zahraei, Polina Stadnikova, and Dietrich Klakow. 2018. Offensive language detection with neural networks for germeval task 2018. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018).*

[von Grunigen et al.2018] Dirk von Grunigen, Ralf Grubenmann, Fernando Benites, Pius Von Daniken, and Mark Cieliebak. 2018. spmmmp at germeval 2018 shared task: Classification of offensive content in tweets using convolutional neural networks and gated recurrent units. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

[Waseem and Hovy2016] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

[Wiedeman et al.2018] Gregor Wiedeman, Eugen Ruppert, Raghav Jindal, and Chris Biemann. 2018. Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

[Wiegand et al.2018] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.