DFKI SLT at GermEval 2021: Multilingual Pre-training and Data Augmentation for the Classification of Toxicity in Social Media Comments

Remi Calizzano
DFKI GmbH
Berlin, Germany
remi.calizzano@dfki.de

Malte Ostendorff
DFKI GmbH
Berlin, Germany
malte.ostendorff@dfki.de

Georg Rehm
DFKI GmbH
Berlin, Germany
georg.rehm@dfki.de

Abstract

We present our submission to the first subtask of GermEval 2021 (classification of German Facebook comments as toxic or not). Binary sequence classification is a standard NLP task with known state-of-the-art methods. Therefore, we focus on data preparation by using two different techniques: task-specific pre-training and data augmentation. First, we pre-train multilingual transformers (XLM-RoBERTa and MT5) on 12 hatespeech detection datasets in nine different languages. In terms of F1, we notice an improvement of 10% on average, using task-specific pre-training. Second, we perform data augmentation by labelling unlabelled comments, taken from Facebook, to increase the size of the training dataset by 79%. Models trained on the augmented training dataset obtain on average +0.0282 (+5%) F1 score compared to models trained on the original training dataset. Finally, the combination of the two techniques allows us to obtain an F1 score of 0.6899 with XLM-RoBERTa and 0.6859 with MT5. The code of the project is available at: https://github.com/airKlizz/germeval2021toxic.

1 Introduction

Toxicity classification, or, more generally, hatespeech detection, has become a highly important topic due to the explosion of social media use. The automation of this task is a challenge for the NLP field with an increasing amount of research on this subject (Schneider et al., 2018; Aluru et al., 2020; Corazza et al., 2020). The GermEval series has already looked into various aspects related to the detection of German language hatespeech with two shared tasks on offensive language identification (Wiegand et al., 2018; Struß et al., 2019). The first subtask of GermEval 2021 follows in these footsteps with the classification of toxic comments.

We want to take advantage of the proliferation of hatespeech datasets for various languages created in the last couple of years. Additionally, in the meantime, a number of multilingual language models have been published (Conneau et al., 2020; Xue et al., 2021; Liu et al., 2020; Lewis et al., 2020) with a high capacity for cross-lingual transfer. We use multilingual models and pre-train them on a multilingual dataset created out of 12 datasets for nine different languages on toxicity and hatespeech detection. We evaluate whether performing this type of pre-training on multilingual models can improve their performance. We assume that the cross-lingual transfer capacity of the multilingual models can be applied to task-specific pre-training and that this will improve final performance on the German-only dataset of the shared task.

Furthermore, we perform data augmentation by labelling unlabeled data, retrieved from Facebook, using one of the multilingual models pre-trained and fine-tuned on the toxicity classification task. As the dataset of the shared task contains only 3244 examples, we hope that extending the number of training examples can improve the overall performance of the models.

In summary, our main contributions are:

- Comparison of the performance of two multilingual models (XLM-RoBERTa and mT5) against a German-specific language model (GBERT) on a German binary classification task with and without task-specific pre-training for multilingual models.
- Evaluation of the models when using data augmentation to increase the size of the dataset used for fine-tuning.

The rest of this article is structured as follows. Section 2 presents our methodology for task-specific pre-training and data augmentation. Section 3 introduces the task as well as the dataset

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

and describes the models and training scenarios. Sections 4 and 5 present and discuss the results obtained in these training scenarios. Concluding remarks are provided in Section 6.

2 Methodology

2.1 Task-specific pre-training

Toxicity or, more generally, hatespeech classification is an NLP task that is supported through multiple datasets in multiple languages. Although the specific task may differ from one dataset to another due to the type of content and annotations used (Bourgonje et al., 2018), the features used to classify sequences are similar.

Pre-training is a technique that often enables state-of-the-art performance in many NLP tasks (Sarlin et al., 2020). Task-specific pre-training has shown its efficiency to produce models that capture task-specific features and that, thus, exhibit better performance (Li et al., 2020).

We want to profit from the many existing hatespeech classification datasets by using these datasets to perform task-specific pre-training.

We adapt task-specific pre-training to toxicity classification by taking 12 toxicity or hatespeech classification datasets and training language models on these datasets before fine-tuning them on the dataset of the shared task (Table 1). Our task-specific pre-training dataset is composed of a total of 105,142 examples in nine different languages.

To take advantage of this task-specific multilingual pre-training, we work with multilingual models. Indeed, these models have already demonstrated their ability to transfer what they have learned in one language into other languages (Hu et al., 2020). In this work, the models will be fine-tuned on the dataset of the shared task which is in German only, however, we assume that the multilingual models can benefit from the task-specific pre-training.

2.2 Data augmentation

In addition to the task specific pre-training, we increase the size of the shared task dataset using data labelling. We use our best performing model and fine-tune on the toxicity classification task of the shared task to label unlabelled Facebook comments we collected from German political talk shows. In total, we collected 5563 Facebook comments added to posts from the pages of ZDF heute[1], Panorama[2], Maischberger[3], and hart aber fair[4]. mT5 is performing better than XLM-RoBERTa on the final toxic classification task when simply using task-specific pre-training and fine-tuning, therefore we use mT5 to compute the probability of a comment to be toxic or not. We only keep the comments classified as toxic or non-toxic with a probability larger than 0.8. Figure 1 shows examples of comments with their toxicity probabilities. This way we label 2044 comments, which we add to the original shared task dataset. Table 2 compares the original dataset with the one we created and also with the augmented dataset which corresponds to the combination of the original dataset and the one we created using data augmentation.

Dataset	Number of examples	Languages
Chung et al. (2019)	7,659	eng, fra, ita
Gao and Huang (2017)	1,528	eng
Wiegand et al. (2018)	5,009	deu
Mandl et al. (2019)	14,336	eng, deu, hin
Ousidhoum et al. (2019)	13,014	ara, eng, fra
de Gibert et al. (2018)	10,944	eng
Davidson et al. (2017)	24,783	eng
Alfina et al. (2017)	713	ind
Ross et al. (2016)	469	deu
Mulki et al. (2019)	5846	apc
Nascimento et al. (2019)	7,672	por
Ibrohim and Budi (2019)	13,169	ind

Table 1: List of all the datasets used for the task-specific pre-training with the number of examples and the languages (code ISO 639-3) for each dataset.

3 Experiments

3.1 Task and dataset

The first subtask of GermEval 2021 is the classification of Facebook comments from German political talk shows with regard to their toxicity. Figure 2 shows two examples. Risch et al. (2021) provide a detailed description of the dataset.

We split the original dataset into a train and an evaluation portion to be able to evaluate our models during training. We use 80% of the original dataset for training and 20% for the evaluation, for which we use precision, recall, and macro-average F1.

[1] https://www.facebook.com/ZDFheute/
[2] https://www.facebook.com/panorama.de
[3] https://www.facebook.com/maischberger
[4] https://www.facebook.com/hartaberfairARD

Comment	Toxicity probability
Hat vermutlich auch überhaupt nichts mit Merkels Desaströser Politik zu tun	0.8790
Frage: Wenn die Tage kürzer werden, das Gehalt aber gleich bleibt, reicht es dann länger?	0.0541
Die Hausärzte bekommen Astra nicht verimpft und die Impfzentren bleiben halb leer. Impfturbo?	0.5627
Na was sind die Bürger erst enttäuscht von euch allen samt dem Gremium....	0.6742

Figure 1: Samples of comments collected on Facebook posts from German political talk shows with their toxicity probability. We only keep the comments classified as toxic or non-toxic with a probability larger than 0.8

	Number of examples		Toxic label	Number of words per comment		
	train	evaluation	ratio	mean	30^{th} pctl	70^{th} pctl
Original GermEval 2021 dataset	2,596	648	0.35	28	11	30
Created dataset	2,044	0	0.49	36	17	39
Augmented dataset	4,640	648	0.40	31	13	34

Table 2: Comparison of the original shared task dataset, the dataset created using data augmentation, and the augmented dataset, i. e., the combination of the other two datasets.

3.2 Models

The task-specific pre-training is based on a multilingual dataset (Section 2.1). We picked two multilingual Transformer models, XLM-RoBERTa and mT5. In addition, we compare multilingual models with the German Transformer based language model GBERT that we evaluate with our data augmentation method.

GBERT GBERT (Chan et al., 2020) is a German language model using the same architecture as BERT (Devlin et al., 2019). GBERT is an encoder-only Transformer model. It was trained using masked language modeling with whole word masking which corresponds to masking all of the tokens corresponding to a word. The pre-training corpus consists of German texts from Wikipedia, Common Crawl (Ortiz Suárez et al., 2019), OPUS (Tiedemann, 2012), and Open Legal Data (Ostendorff et al., 2020). GBERT outperforms the state-of-the-art for the GermEval 2018 hatespeech detection task and the GermEval 2014 NER task (Chan et al., 2020). We use the GBERT Base version.

XLM-RoBERTa XLM-RoBERTa (Conneau et al., 2020) is the multilingual version of RoBERTa (Liu et al., 2019). It was trained on the Common Crawl corpus in 100 languages using masked language modeling. We choose XLM-RoBERTa instead of Multilingual BERT[5] because XLM-RoBERTa outperforms Multilingual BERT on a variety of cross-lingual benchmarks (Conneau et al., 2020). We use the Base version of XLM-RoBERTa.

mT5 mT5 (Xue et al., 2021) is a multilingual variant of T5 (Raffel et al., 2020) covering 101 languages. It uses the same architecture as T5, an encoder-decoder Transformer model. Being a text-to-text model, we transform the binary classification task into a text generation task where we train mT5 to generate "neutral" when the input label corresponds to a non-toxic comment and "toxic" when the input label is toxic. We also add the task prefix "speech review" at the beginning of each input sequence. As T5, mT5 exists in five sizes: Small, Base, Large, XL, XXL. The XXL version of mT5 performs better than other multilingual models such as XLM-RoBERTa on many multilingual benchmarks, however, due to computational limits, we use the mT5 Base version that produces results comparable to XLM-RoBERTa (Xue et al., 2021).

3.3 Training scenarios

To evaluate the benefit of the task-specific pre-training and data augmentation, we train the models in four different scenarios.

Fine-tuning only We first fine-tune the three models on the original dataset of the shared task. These models are used as baselines to evaluate the two methodologies we propose.

With task-specific pre-training In this scenario, we pre-train mT5 and XLM-RoBERTa on the task-specific pre-training dataset (Section 2.1). The task-specific pre-training consists of training the models with the same objective as the fine-tuning task

[5] https://github.com/google-research/bert/blob/master/multilingual.md

Comment	Toxicity
Die SPD, Verbrecher,die haben Angst vor den Wahlen in den neuen Bundesländern,weg mit Euch.	1
Ich schmeiß mich weg... 800 Euro sollen für ein ””vernünftiges”” Leben ausreichen?	0

Figure 2: Two comments from the original GermEval21 shared task dataset with their toxicity labels.

which is the classification of toxic comments. As the result of the combination of those datasets is not balanced, we randomly remove non-toxic samples to arrive at the same number of toxic and non-toxic samples. Afterwards, we fine-tune the task-specific pre-trained models as in the first scenario.

With data augmentation This scenario corresponds to the first one except we use the augmented dataset instead of the original shared task dataset. The augmented dataset combines the original and one additional dataset (Table 2).

With task-specific pre-training and data augmentation This scenario combines the second and third scenario. We fine-tune the task-specific pre-trained models on the augmented dataset.

We use the HuggingFace Transformers library (Wolf et al., 2020) to train the models. GBERT and XLM-RoBERTa are trained using the hyperparameter search method[6] with Optuna as the optimization framework[7], the maximization of the F1 metric as computing objective, and a number of trials equals to 10. As mT5 requires more training time, we do not use hyperparameter search for mT5 but fixed parameters that we found to be the best. We use a learning rate of 5^{-5}, a batch size of 16, and we train mT5 for 3 epochs. In the end we select the best model with regard to the F1 score.

To deal with the imbalanced training dataset, we use class weights for GBERT and XLM-RoBERTa and oversample the dataset for mT5.

4 Results

We evaluate the models on the test dataset provided by the organizers of the shared task after the training phase and the submissions (see Table 3).

First, adding task-specific pre-training and/or using data augmentation improves the results for both XLM-RoBERTa and mT5. Training with task-specific pre-training and data augmentation improves the F1 score by 0.0490 (+8%) for XLM-RoBERTa and by 0.0836 (+14%) for mT5. GBERT also produces slightly better results, the F1 score improves by 0.0066 (+1%), when using the augmented dataset for fine-tuning.

Model	F1	Precision	Recall
Fine-tuning only			
GBERT	0.6663	0.6437	0.6906
XLM-RoBERTa	0.6409	0.6373	0.6445
mT5	0.6023	0.5995	0.6052
With task-specific pre-training			
XLM-RoBERTa	0.6785	0.6851	0.6720
mT5	0.6799	0.6840	0.6759
With data augmentation			
GBERT*	0.6729	0.6724	0.6734
XLM-RoBERTa	0.6680	0.6720	0.6639
mT5	0.6533	0.6541	0.6526
With task-specific pre-training and data augmentation			
XLM-RoBERTa*	**0.6899**	**0.6900**	**0.6898**
mT5*	**0.6859**	**0.6899**	**0.6818**

Table 3: F1, recall and precision results of each model on the test dataset of the shared task for each training scenario. * *models used for our submissions. Results slightly differ from the submissions because we retrained all the models for the paper.*

Second, for the models fine-tuned only on the original dataset, mT5 obtains the worst results with an F1 score of 0.6023, followed by XLM-RoBERTa with 0.6409, and GBERT with 0.6663. The ranking is the same for the models fine-tuned on the augmented dataset but with a smaller gap between scores. F1 scores for mT5, XLM-RoBERTa and GBERT are 0.6533, 0.6680 and 0.6729.

Third, despite mT5 performing worse that XLM-RoBERTa by 0.0386 when fine-tuned on the original dataset, the results with task-specific pre-training and data augmentation of the two models are very similar with a difference between F1 scores lower than 0.1%. This correlates with the fact that the task-specific pre-training particularly improves the results of mT5 with an increase of 0.0776 (+13%) of the F1 score compared to an increase of 0.0376 (+6%) for XLM-RoBERTa.

Overall, XLM-RoBERTa and mT5 with task-specific pre-training and data augmentation are the

[6] https://huggingface.co/transformers/main_classes/trainer.html#transformers.Trainer.hyperparameter_search

[7] https://optuna.org

models that obtain the best F1 scores with 0.6899 and 0.6859, respectively.

5 Discussion

In the two scenarios where only German data is used (*Fine-tuning only* and *With data augmentation*), GBERT performs better than XLM-RoBERTa and mT5. This is easily explained by the fact that GBERT was pre-trained only on German data, in contrast to mT5 and XLM-RoBERTa. However, the small difference in F1 scores with the use of the augmented dataset (*With data augmentation*) implies that with more data, multilingual models can perform as well as monolingual models. Additionally, we see that the task-specific pre-training of multilingual models on a multilingual dataset compensates for the poorer performance of mT5 and XLM-RoBERTa when trained on a German only dataset compared to GBERT. It is interesting to note that the task-specific pre-training of mT5 and XLM-RoBERTa on a multilingual dataset allows them to perform better than GBERT. The fact that multilingual models can benefit from hatespeech classification datasets in other languages allows them to perform better than the German-only model. It is also important to notice that XLM-RoBERTa and mT5 use more recent architectures and/or pre-training methods than GBERT. It may also partly explain that GBERT's results are worse than those of XLM-RoBERTa and mT5.

Moreover, as noted in Section 4, XLM-RoBERTa does not benefit from the task-specific pre-training as much as mT5. Our hypothesis is that having less trainable parameters, XLM-RoBERTa (270M parameters) does not have as much capacity as mT5 (580M parameters) to benefit from all the examples on which the models are pre-trained. The number of parameters of the models is an important aspect to take into consideration when doing pre-training in general, and we observe this again in our experiments with task-specific pre-training.

6 Conclusion

We describe the methods used for our submissions to the GermEval 2021 toxic comment classification task. Specifically, we can benefit from hatespeech detection datasets in other languages to improve the performance of multilingual models through task-specific pre-training. With this method, multilingual models (XLM-RoBERTa and mT5) perform even better, $+0.0576$ ($+10\%$) in average in terms of F1, than GBERT, a German-specific language model. We show that by increasing the shared task dataset by automatically labeling additional comments from Facebook, we are able to improve the results of the three models we evaluated (GBERT, XLM-RoBERTa, mT5) by 5% in average.

We have shown that multilingual models can perform as well or even better than monolingual models by performing task-specific multilingual pre-training. This particularly applies to tasks for which many datasets are available in languages different from the dataset used for fine-tuning and where the fine-tuning dataset is relatively small (less than 10,000 samples) as is the case of the German toxic comment classification task.

In addition, multilingual models have some other advantages. First, in a production setting, it might not be feasible to deploy multiple monolingual models due to resource constraints. Replacing multiple monolingual models with a single multilingual model can be a solution. Second, multilingual models, due to their cross-lingual transfer capacity, can be used in a language other than the language of the training dataset. This allows the creation of models for languages for which obtaining training data can be difficult.

Acknowledgments

The research presented in this paper is funded by the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (http://qurator.ai) (Unternehmen Region, Wachstumskern, grant no. 03WKDA1A). In addition, the authors would like to thank Melina Plakidis for her contribution regarding the data labelling.

References

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *ArXiv*, abs/2004.06465.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2018. Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication. In *Language Technologies*

for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 180–191, Cham, Switzerland. Gesellschaft fur Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020. Task-specific objectives of pre-trained language models for dialogue adaptation. *ArXiv*, abs/2009.04984.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

Gabriel Nascimento, Flavio Carvalho, Alexandre Martins da Cunha, Carlos Roberto Viana, and Gustavo Paiva Guedes. 2019. Hate speech detection using brazilian imageboards. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, WebMedia '19, page 325–328, New York, NY, USA. Association for Computing Machinery.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 385–388, New York, NY, USA. Association for Computing Machinery.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Julian Moreno Schneider, Roland Roller, Peter Bourgonje, Stefanie Hegele, and Georg Rehm. 2018. Towards the Automatic Classification of Offensive Language and Related Phenomena in German Tweets. In *Proceedings of the GermEval Workshop 2018 – Shared Task on the Identification of Offensive Language*, pages 95–103, Vienna, Austria. 21 September 2018.

Julia Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Zurich Open Repository and Archive*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.