DeTox at GermEval 2021: Toxic Comment Classification

Mina Schütz[1], Christoph Demus[2], Jonas Pitz[1], Nadine Probol[1], Melanie Siegel[1], Dirk Labudde[2]

[1] Darmstadt University of Applied Sciences
Max-Planck-Straße 2, 64807 Dieburg
{mina.schuetz, melanie.siegel}@h-da.de
{jonas.pitz, nadine.probol}@stud.h-da.de

[2] Fraunhofer Institute for Secure Information Technology
Rheinstraße 75, 64295 Darmstadt
{christoph.demus, dirk.labudde}@sit.fraunhofer.de

Abstract

In this work, we present our approaches on the toxic comment classification task (subtask 1) of the GermEval 2021 Shared Task. For this binary task, we propose three models: a German BERT transformer model; a multilayer perceptron, which was first trained in parallel on textual input and 14 additional linguistic features and then concatenated in an additional layer; and a multilayer perceptron with both feature types as input. We enhanced our pre-trained transformer model by re-training it with over 1 million tweets and fine-tuned it on two additional German datasets of similar tasks. The embeddings of the final fine-tuned German BERT were taken as the textual input features for our neural networks. Our best models on the validation data were both neural networks, however our enhanced German BERT gained with a F1-score = 0.5895 a higher prediction on the test data.

1 Introduction

In recent years social media platforms became a popular medium to discuss all kinds of topics with people around the world. Also shops, companies, TV-shows and many more use social media to present their content to followers and discuss it with them. As it is possible to interact almost anonymously on the internet, such social media pages are often confronted with the problem of hate speech and toxic comments targeting single persons or whole groups (Watanabe et al., 2018). Although hate speech detection has been a top research topic for several years, there exists no satisfactory solution yet (Struß et al., 2019). The GermEval Shared Task 2021 (Risch et al., 2021) addresses this topic - especially the side of social media moderators that are responsible to filter such comments - in this years challenge with the following three tasks, where we participate in subtask 1:

- Subtask 1: toxic comment classification
- Subtask 2: engaging comment classification
- Subtask 3: fact-claiming

Over the last years transformer (Vaswani et al., 2017) models like BERT (Bidirectional Encoder Representations with Transformers) (Devlin et al., 2019) became state-of-the-art for many natural language processing (NLP) tasks and regularly outperformed traditional machine learning models and neural networks (Zampieri et al., 2020; Kumar et al., 2020). Nevertheless, the GermEval Shared Task 2019 showed that traditional machine learning methods can still achieve comparable results to the transformer models if the features are well chosen (Struß et al., 2019).

Therefore, we decided to experiment with standard supervised machine learning models and neural networks, different word embeddings, and pre-trained transformer models. We then chose our best performing transformer model, enhanced it with re-training on extracted tweets in German, and fine-tuned it with additional datasets. The extracted word embeddings by our transformer model were used as an textual input for our neural network architectures besides additional features.

Our presented work is structured as follows: Section 2 gives an overview of related work. In Section 3 we describe the GermEval 2021 data and the additional data we used for our final models. In Section 4 the feature extraction, the baseline and the final models are described. In Section 5, we show our final results and discuss our models.

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

2 Related Work

Toxic speech can be defined as a combination of hate speech and offensive language (D'Sa et al., 2020) or a type of aggressive writing style (Maslej-Krešňáková et al., 2020). Many recent research uses deep neural networks for such detection tasks in social media content (Georgakopoulos et al., 2018; van Aken et al., 2018). There has also been some research with transformer models, especially for English social media content. Maslej-Krešňáková et al. (2020) compared multiple transformers and neural networks for the classification of toxic content with different types of preprocessing steps, focussing on word embeddings. However, some related work to our modelling approach has been done by researchers in similar content detection tasks on social media.

Sohn and Lee (2019) used, in their study on hate speech detection with transformer models, a similar approach to our proposed models, after they fine-tuned a multi-channel BERT model: they applied a dropout on the [CLS] token of BERT and added a feed forward layer before the softmax output and calculated the weighted sum of three transformers instead of only one. The [CLS] token is the final hidden vector of BERT used for classification, however it can also be extracted for the models embeddings (Devlin et al., 2019). This was also done in (Rodríguez-Sánchez et al., 2020) for the task of automatic sexism classification, where the authors added features with a feed forward layer on top, however this did not improve their results. They also - in comparison to our concatenation strategy for our multilayer perceptron - created a Bi-LSTM (Bidirectional Long-Short-Term-Memory), where they concatenated the additional extracted features (in this case user and network information) after going through several layers of the neural network with only using textual input. Their work showed that using pre-trained embeddings for neural networks pushes the final classification by 3% (Rodríguez-Sánchez et al., 2020).

The study of Zhao et al. (2021) found that using pre-trained models as an input for neural networks leads to better results than using complex deep neural networks or transformers as a stand-alone architecture. Comparingly, another approach by D'Sa et al. (2020) on hate speech detection analyzed FastText (Bojanowski et al., 2017) and BERT embeddings and used them as the input for deep neural networks without any additional feature extraction. They found that fine-tuning transformers without a neural network layer performs better.

Those studies show that combining transformers that are fine-tuned for a specific NLP task with neural networks is a promising approach to create better models for predicting toxic comments. Since transformers are usually only used for training on the textual input, the feed forward layers can be concatenated with more extracted features.

	Toxic	Not Toxic	Total
Train	1122 (35.6%)	2122 (64.4%)	3244
Test	350 (37.1%)	594 (62.9%)	944
Total	1472	2716	4188

Table 1: Class distribution for subtask 1 of the GermEval 2021 dataset. Percentages show the proportion of toxic and non-toxic comments in the training and test set.

3 Data

In this section we describe the GermEval 2021 Shared Task dataset as well as the supplementary datasets that we used for fine-tuning our model.

3.1 GermEval 2021 Data

The dataset for the GermEval 2021 Shared Task contains 3244 user comments from the Facebook discussion page of a German news broadcast within the first half of 2019. The comments were anonymized and cleared of any references to the show, moderators and users. The dataset was provided with manual annotated labels for each of the subtasks. Table 1 shows that 35.6% of all comments are labeled as *Toxic* for subtask 1 while 64.4% are labeled as *Not Toxic*.

3.2 Additional Datasets

Augmentation allows a transformer model to be fine-tuned with additional labeled data (Schütz et al., 2021). In order to augment the GermEval 2021 training data we identified two German datasets that were labeled for hateful or offensive comment classification and shared a similar domain. We assumed that the tasks of identifying hateful and offensive comments should be similar to the task of identifying toxic comments.

- **GermEval 2019:** Task 2 of GermEval 2019 was a shared task on the identification and categorization of offensive language (Struß et al., 2019). For subtask 1 of this shared task a total

of 7025 tweets were collected and labeled as either *OFFENSE* or *OTHER* with 32.1% of the tweets being labeled the former. The label *OFFENSE* was given to any comment that was deemed abusive, insulting and/or profane. Comparably to what we would expect from comments about a daily talk show the tweets in this dataset were chosen to cover a broad range of topics.

- **HASOC 2019:** HASOC (Hate Speech and Offensive Content Identification in Indo-European Languages) 2019 was a shared task comparable to GermEval Task 2 but with the addition of providing 3 separate datasets for German, English and Hindi (Mandl et al., 2019). The German dataset contains a total of 4669 tweets and Facebook posts collected by searching for offensive keywords and hashtags. 11.6% of the entries for subtask 1 are labeled as *HOF* while the rest is labeled as *NOT*. The categories *HOF* and *NOT* directly correspond to the categories *OFFENSE* and *OTHER* from Task 2 of GermEval 2019.

3.3 German Tweet Corpus:

For several unsupervised training steps in our experiments we also collected a total of unlabeled 1,156,458 German tweets of the first half year of 2019 via the Twitter API. Mainly, we focused on general tweets in German, as well as tweets from the Twitter pages of German talk shows and other socially critical TV-formats: "Hart aber Fair", "Maybrit Illner", "Anne Will", "Markus Lanz", "ZDF heute-show" and "Maischberger". With this extra data we expected to enhance the predictions of our models, since the dataset hopefully contains tweets with a similar writing style and domain-specific politically discussed content by that time period.

4 Methodology

In this section the feature extraction methods as well as the baseline we used for comparison, the conducted preprocessing steps, and final models are described. Our baseline models include different combinations.

4.1 Feature Extraction

For training some of our models, we used several features as listed in Table 2. It has been shown that adding more specific features about the writing style of social media entries helps to improve the results of similar NLP tasks, such as hate speech and disinformation detection (Robinson et al., 2018; Volkova and Jang, 2018). For toxic comment classification we considered the word count for each input and extracted the number of punctuation, exclamation, and question marks and their relation to the total number of words per comment. For some features we used additional non-public word lists and libraries and cross-checked them for each entry in the dataset:

Feature	Toxic	Not Toxic
word count	201	179
punctuation count	7.41	6.84
exclamation count	0.69	0.31
question mark count	0.48	0.36
word punctuation ratio	0.0111	0.0138
word exclamation ratio	0.0027	0.0021
word question mark ratio	0.0020	0.0030
hate word count	0.32	0.24
hate word count ratio	0.0017	0.0014
character capslock ratio	0.0306	0.0168
sentiment	-0.0147	-0.0080
emoji count	0.49	0.13
emoji sentiment	0.0424	0.0191
word emoji ratio	0.0457	0.0227

Table 2: Extracted features and their mean values in toxic and non-toxic comments.

- "Sentiment" features: list of 9,382 words and their sentiment values
- "Hate" features: list of 3,550 words

Lastly, we counted the number of emojis per comment, determined the emoji word ratio and used the emosent[1] library to compute the average sentiment over all emojis in a comment.

We computed the mean values of each feature for both classes and found some significant differences between both categories: for example toxic comments are 22 words longer on average. Besides the length, there is a notable difference in the number of exclamation marks and emojis between toxic and not toxic comments. Contrary to the expectations the sentiment of the comments is in both cases slightly negative and does only differ by 0.0067 on a scale from -1 (most negative) to +1

[1] https://pypi.org/project/emosent-py/

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

(most positive). Nevertheless, we used all of the extracted features for our experiments.

4.2 Baseline

For our baseline we used a Support Vector Machine (SVM) and a sequential neural network (multilayer perceptron, MLP). Additionally, a Robust Soft Learning Vector Quantization (RSLVQ) model was trained and evaluated. RSLVQ is an adaption of the LVQ Model introduced by Kohonen (1997). In these models, class regions are defined by prototype vectors in the vector space, where each class has one or more prototype vectors. In contrast to the basic LVQ, which is a heuristic, RSLVQ can be mathematically verified (Schneider et al., 2009). Additionally, we tested three pre-trained transformer (Vaswani et al., 2017) models with only using the provided training set by the GermEval 21.

4.2.1 Preprocessing

Multiple preprocessing steps were applied to the SVM and RSLVQ, and the comments were vectorized. The steps included tokenization, stop word and punctuation removal and lemmatization. Hashtags and mentions were preserved in the data, only the characters "#" and "@" were removed. Afterwards 200-dimensional FastText word embeddings were trained on the preprocessed training dataset, on our self collected German Tweet corpus, and on the additional data. For the word embeddings, a skip-gram model with a window-size of 5 and a minimum word occurrence of 3 was used. All the word-vectors of every comment were averaged to receive a document vector.

Additionally, a feature vector for every comment, including the features mentioned in Table 2, was created from the original (not preprocessed) data and concatenated with the document vector.

In contrast, we did not preprocess the data for the transformer models, since those models capture the context of a sentence and use a already specialized built-in tokenizer (Devlin et al., 2019). All of our baseline models were evaluated on a stratified 90% training and 10% validation split.

4.2.2 Experiments

The SVM was trained on the training split using a Radial Basis Function (RBF) and a linear kernel. The best results were achieved with the RBF-kernel. In the RSLVQ model the number of prototypes per class was varied having the best results with two prototypes per class. Already pre-trained FastText embeddings were used as an input for the MLP, where we concatenated the extracted features with the textual input during training (MLP-C) and before (MLP-B). Even though the precision and recall were higher compared to the other models, we found inconsistency in the evaluation plots of the metrics of both models - and due to a high loss during validation, it seemed that both MLPs were overfitting.

Model	Val Pre	Val Rec	Val F1
SVM*	0.57	0.63	0.60
RSLVQ*	0.70	0.43	0.54
MLP-C*	0.65	0.99	0.78
MLP-B*	0.66	0.98	0.79
BERT	0.66	0.65	0.64
DistilBERT	0.67	0.67	0.66
XLM-R	0.71	0.68	0.67

Table 3: Baseline results on the validation split of the GermEval 2021 training data.
*Additional German tweets used for word embeddings.

Finally, we fine-tuned a German BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) model (bert-base-german-cased (Chan et al.), distilbert-base-german-cased (Chaumond)) provided by the HuggingFace library (Wolf et al., 2020) for 10 epochs, a batch size of 16, a learning rate of 2e-5, Adam (Kingma and Ba, 2015) as an optimizer and a maximum sequence length of 256. The multi-lingual transformer XLM-R (Conneau et al., 2019) was fine-tuned with the same parameters, except a learning rate of 1e-5 instead.

4.3 Models

In total we submitted three different models for each run as shown in Figure 1.

- **Transformer (TAB):** We decided to enhance our best transformer model from our baseline by using the additional German tweets for re-training. This has been shown to help boost the classification accuracy as shown in (Schütz et al., 2021). Re-training means that the pre-trained model is further trained in an unsupervised manner, before fine-tuning it for the NLP downstream task. We chose to re-train with the the german-bert-base-cased model for 5 epochs, with a batch size of 32 and a learning rate of 2e-5. Afterwards, we fine-tuned our re-trained German-BERT model

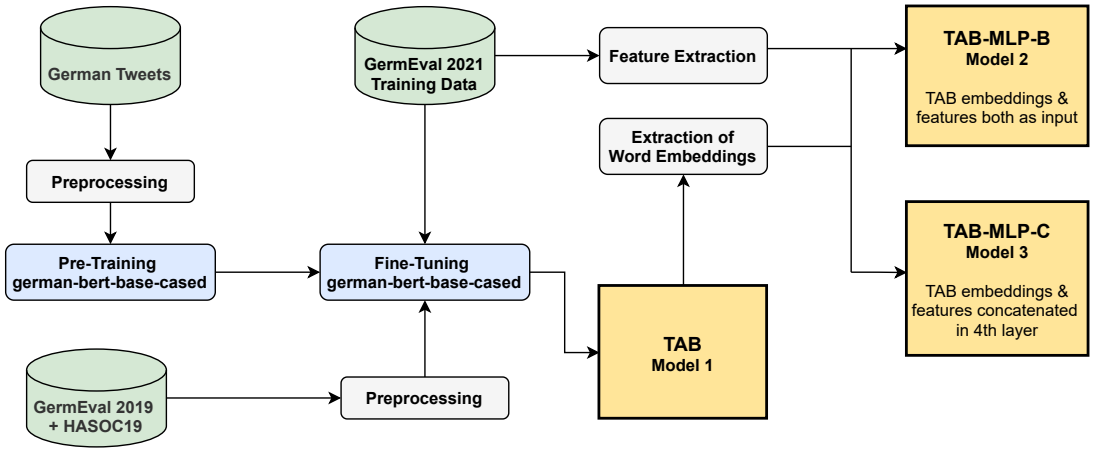

Figure 1: Experimental setup for training our submitted models (Green: datasets; grey: processing steps; blue: transformer re-training & fine-tuning steps; yellow: final models).

on the GermEval 2021 training data, as well as the additional datasets (GermEval 2019 & HASOC 2019). The augmented dataset contained a total of 24,304 comments, where 5,414 we set as toxic and 18,890 as not toxic as described in section 3. However, we added one more preprocessing step, compared to the transformer baselines, for pre-training and fine-tuning our model, since the authors of the GermEval 2021 changed every username in the comments to "@USER". We applied this to the additional German tweets as well as to the GermEval 2019 and HASOC 2019 datasets to align all texts. For the evaluation of our model, we used 10% of the GermEval 2021 training dataset. Our final transformer model, called TAB (tweets-and-Additional-Datasets-BERT) was trained on this augmented data for 10 epochs, a batch size of 16, a learning rate of 2e-5, Adam as an optimizer, and a maximum sequence length of 256.

- **Multi-Layer Perceptron (TAB-MLP):** For our second and third run, we used the MLP model we created for the baseline. Its architecture consists of 5 dense layers, a dropout of 0.2, ReLU (Rectified Linear Unit) as an activation function and sigmoid for our final classification layer. Since the FastText embeddings seemed to overfit the model, we extracted the already fine-tuned word embeddings of the TAB model via the [CLS] token of each input. Lastly, the additional extracted features were normalized and used for two different training strategies:

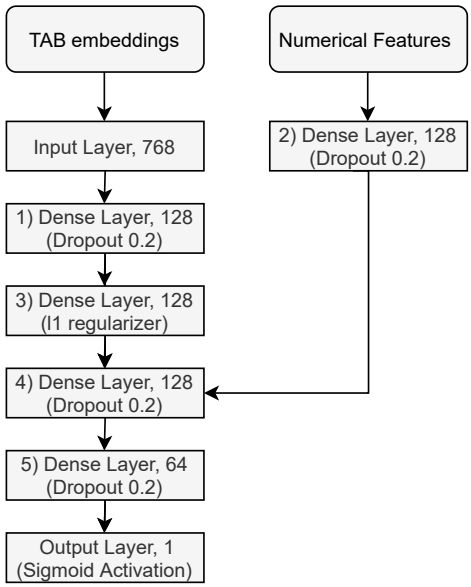

Figure 2: Architecture of TAB-MLP-C.

- *TAB-MLP-B:* the model was fed with the text input as well as the features combined as one input vector for training.
- *TAB-MLP-C:* the model was trained on the textual input for 3 layers, the numerical features for 1 layer, and then concatenated in the 4th layer as shown in Figure 2.

Both models were trained for 25 epochs, a batch size of 32, a learning rate of 1e-2, and Stochastic Gradient Descent (SGD) as an optimizer. After plotting the curves of the evaluation metrics and comparing them with the FastText embeddings (Table 3) we found that the MLP did not seem to

Model	Run	Val Precision	Val Recall	Val F1	T Precision	T Recall	T F1
TAB	1	0.74	0.68	0.68	**0.6306**	**0.5535**	**0.5895**
TAB-MLP-C	2	**0.67**	**0.94**	**0.78**	0.3622	0.3597	0.3572
TAB-MLP-B	3	0.65	0.98	0.78	0.3854	0.3771	0.3812

Table 4: Results of our proposed models on the validation (Val) split of the training set and the test data (T).

Model	TP	TN	FP	FN
TAB	61	554	40	289
TAB-MLP-B	144	180	414	206
TAB-MLP-C	122	241	353	228

Table 5: Confusion matrix for each of our submitted models (TP: true positives, TN: true negatives, FP: false positives, FN: false negatives).

overfit with the already pre-trained TAB embeddings. Since we used a sigmoid activation function in our classification layer, we set a threshold for the predictions on the test set at 0.7, after calculating the mean and median value for each of our neural networks.

5 Results and Discussion

All of our models were evaluated with precision, recall, and a macro-averaged F1-score as shown in Table 4. The final results on the test data show that the transformer model gained by far the best results with its F1-score of 0.5895, even if it is still not as high as the value we expected after our training validation. Our neural networks TAB-MLP-B and TAB-MLP-C performed significantly worse on the test data, especially with regard to their high F1-score on the validation split.

Therefore, we explored whether we set the threshold too high for our predictions on the test data. Even though we experimented with setting the threshold to different values, we found that the predictions did not improve significantly (only ≈ 0.01), which shows that the neural networks probably overfitted on one class. We suspect this is also the reason for the very high validation recall in comparison to the precision. We plotted the confusion matrix for each model, shown in Table 5, which shows that both neural networks had a high count of false positives. In contrast to that, TAB had an issue with the false negatives. Therefore, we conclude several possible reasons why our neural networks did not perform well on the test set:

- the size of the dense layers, type of activation function and dropout have to be adjusted.
- the additional features have no positive impact on the models.
- another embedding strategy for the transformer models carries more information than the extraction of the [CLS] token. A possible solution could be a concatenation of a number of hidden layer outputs.

6 Conclusion

In this work we presented our submitted models for the GermEval Shared Task 2021 on toxic comment classification. We decided to combine standard supervised methods with transformers and textual features, and to enhance the models with additional training data.

Our best model was a German BERT that was re-trained on over 1.5 million additional German tweets from the first half year of 2019 and fine-tuned with two augmented datasets from similar tasks, such as hate speech and offensive language detection, as well as the GermEval 2021 training data. Even though our two multilayer perceptrons - which were trained on the extracted word embeddings by our transformer - showed better evaluation results during validation, our BERT model still had a more robust prediction on the test set. For future work, we will further explore the combination of sequential neural networks and word embeddings by transformers and test several extraction and concatenation strategies.

7 Acknowledgements

This work is enhanced by the Darmstadt University of Applied Sciences in collaboration with the Fraunhofer Institute for Secure Information Technology. The Darmstadt University supported this work with the research in Information Science (https://sis.h-da.de/). Additionally, this contribution has been funded by the project ”DeTox” (Cybersecurity research funding of the Hessian Ministry of the Interior and Sports).

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. HuggingFace German BERT. Accessed: 2021-06-10.

Julien Chaumond. HuggingFace German DistilBERT. Accessed: 2021-06-10.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. BERT and fastText embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, pages 1–5.

Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, New York, NY, USA. Association for Computing Machinery.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Teuvo Kohonen. 1997. Learning vector quantization. In *Self-Organizing Maps*, pages 203–217. Springer Berlin Heidelberg.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

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*, pages 14–17.

Viera Maslej-Krešňáková, Martin Sarnovský, Peter Butka, and Kristína Machová. 2020. Comparison of deep learning models and various text preprocessing techniques for the toxic comments classification. *Applied Sciences*, 10(23).

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.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on twitter: Feature engineering v.s. feature selection. In *The Semantic Web: ESWC 2018 Satellite Events*, pages 46–49, Cham. Springer International Publishing.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 8:219563–219576.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Petra Schneider, Michael Biehl, and Barbara Hammer. 2009. Distance learning in discriminative vector quantization. *Neural Computation*, 21(10):2942–2969.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2021. Automatic sexism detection with multilingual transformer models. *arXiv preprint arXiv:2106.04908*.

Hajung Sohn and Hyunju Lee. 2019. MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In

Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), pages 354–365, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 575–583, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. *A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification*, page 500–507. Association for Computing Machinery, New York, NY, USA.