# TUW-Inf at GermEval2021: Rule-based and Hybrid Methods for Detecting Toxic, Engaging, and Fact-Claiming Comments

**Kinga Gémes**
TU Wien
kinga.gemes@tuwien.ac.at

**Gábor Recski**
TU Wien
gabor.recski@tuwien.ac.at

## Abstract

This paper describes our methods submitted for the GermEval 2021 shared task on identifying toxic, engaging and fact-claiming comments in social media texts (Risch et al., 2021). We explore simple strategies for semi-automatic generation of rule-based systems with high precision and low recall, and use them to achieve slight overall improvements over a standard BERT-based classifier.

## 1 Introduction

We present our systems submitted to the GermEval 2021 shared task on identifying toxic, engaging and fact-claiming comments in social media texts (Risch et al., 2021). We focus on strategies for building simple rule-based systems that are both explainable and customizable for end users. We also train a simple BERT-based classifier for comparison, and to evaluate its performance when combined with our high-precision rule-based systems. After a short overview in Section 2 of the task and the datasets used we describe our methods for creating rule-based systems in Section 4 and the BERT-based baseline system in Section 3. Section 5 describes how these systems were combined into simple ensemble models, Section 6 presents quantitative results on the 2021 test set, and a manual qualitative analysis on a sample of our output for Subtask 1 is provided in Section 7. All systems described in this paper are publicly available under an MIT license from the repository `tuw-inf-germeval2021`[1], along with instructions for reproducing our results.

[1] https://github.com/GKingA/tuw-inf-germeval2021

## 2 Task and datasets

The dataset of the 2021 shared task contains 3,244 comments from the Facebook page of a German news broadcast, from discussions between February and July 2019, manually annotated for three categories corresponding to the three subtasks: whether a comment is toxic, engaging, and/or fact-claiming. Definitions for each category and a detailed description of the annotation process are given in the task overview paper (Risch et al., 2021). For developing our rule-based system for toxicity detection we also used a corpus of annotated tweets from Germeval challenges of previous years (Wiegand et al., 2018; Struß et al., 2019), the 2018 and 2019 datasets contain nearly 11,000 German tweets.

Comments in the 2021 dataset were parts of discussion threads related to individual news items. The dataset does not contain such threads, only individual comments, and this is also how they were presented to annotators. However, some fragments of the initial posts ('teaser texts') were made available to annotators as context, but were not included in the dataset because of privacy concerns (Wilms, 2021). This means that in some cases our models may not have had access to the full information that led annotators to their decisions. Some possible examples will be presented in the manual analysis in Section 7. When experimenting with our methods, we split the 3,244 comments of the training dataset into two parts, training our ML models and developing our rules using only 2,434 comments (75 %) and validating our approach on the remaining 811 (25 %).

Some entities in the dataset have been anonymized by the organizers, introducing



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

placeholders such as *@USER*, *@MEDIUM*, and *@MODERATOR*. In addition we also masked URLs, currency symbols, and numbers. For our BERT-based experiments we also replaced emoticons with their German textual representations, using the `emoji` library[2]. A German dictionary has been added to this library only days before the submission deadline. In our submissions we use our own dictionary[3], created from the English resource using the Google Translate API via the `translate` Python library[4].

## 3 BERT-based classification

Language models based on the Transformer architecture (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) provide the basis of strong baseline systems across a wide range of tasks in natural language processing, and some of the top-performing systems in the 2019 GermEval challenge also use BERT (Paraschiv and Cercel, 2019; Graf and Salini, 2019). For our experiments we used the model `bert-base-german-cased`[5] a publicly available BERT model trained only on German data. For each subtask we trained a neural network with a single linear classification layer on top of BERT. Metaparameters were set based on performance on the validation set. We used Adam optimizer with a weight decay value of $10^{-5}$ and initial learning rate of $10^{-5}$. We set batch size to 8 and trained each model for 10 epochs and determined the optimal number of iterations based on either F-score or precision on the validation set (see Section 5 for details). For the final submissions we trained on the full training set (including the validation set).

## 4 Rule-based methods

We explore simple strategies for both manual and semi-automatic generation of lists of words and phrases that can be used in rule-based systems that consider a comment toxic if and only if it contains any of the words or phrases in a list. Our goal is to facilitate the rapid creation of such simple rule-based systems because they are often preferred in real-world applications due to their fully transparent and explainable nature. Any decision made by such a system, whether true or false, positive or negative, can be directly attributed to one or more terms in the input or to the fact that no terms in the input are present in the list of key terms. This offers straightforward ways for users to update the rules in a way that changes a particular decision, by removing keyphrases causing false positives or adding them to fix false negatives. Whether or not this process is actually beneficial for the overall accuracy of a model, it is in line with common business needs, most typically with the common experience that once users have reported an error, they expect it to be corrected. The experiments in this section and the qualitative analysis in Section 7 were performed only for the toxicity detection task.

For the toxicity detection task we experimented with simple strategies for automatic bootstrapping of keyword lists, which are then reviewed and corrected manually. The method involves extracting simple patterns from comments in the training data and ranking them according to their potential as rules, i.e. looking for patterns that in themselves have a very high precision as predictors of toxicity. We searched for patterns characteristic of comments labeled as toxic in the form of a few simple feature types, including unigrams and bigrams of words or lemmas, with or without part-of-speech tag for potential disambiguation purposes. We also tried limiting the space of unigram features to nouns only, or nouns and adjectives only. For tokenization, lemmatization, and part-of-speech tagging we use the `Stanza` library (Qi et al., 2020) using the `gsd` German model with `resources_version` 1.2.0. We achieved best results when limiting our search to word unigrams only and ranking them separately for nouns (including proper nouns) and for all other parts-of-speech.

To extract patterns with a high potential as rules we experimented with simple scoring schemes for ranking all features based on the number of true and false positives they would contribute if used as strict rules, i.e. the number of positive and negative examples containing them as patterns. To assess the efficiency of these strategies, i.e. whether the patterns they extract are generally good rule candidates that can be edited into curated lists, we observed their behaviour in the portion of the dataset

[2] https://github.com/carpedm20/emoji
[3] https://github.com/GKingA/emoji
[4] https://pypi.org/project/translate/
[5] https://deepset.ai/german-bert

used for training the ML models. The validation set was only used for infrequent overall quantitative evaluations and not for observing patterns, since for the purposes of manual rule creation this would have meant using the validation data for training. The strategy we found most effective was to consider patterns with at least 5 occurrences in the training dataset and rank them with the scoring scheme $TP - 100 \cdot FP$, where $TP$ and $FP$ are the number of true and false positives detected by that pattern.

Lists created this way require manual editing so as not to introduce artefacts. For example, the top words in the training dataset for this year's Germeval task contain words like *Hamburg* and *fleissig* 'hard-working' because these words happened to occur in several comments labelled as toxic but none of the non-toxic ones, thereby getting ranked just as high as *Dummheit* 'stupidity' and *Bullshit*, terms that we actually want to keep for the edited list. The majority of good patterns comes from the larger toxicity dataset available to us, the 2018 and 2019 Germeval training datasets (Wiegand et al., 2018; Struß et al., 2019). While in the smaller 2021 dataset the top-ranked patterns occurred in no more than 3 or 4 positive examples of toxicity, the combined training datasets of previous years allowed us to find patterns with 15-25 positive examples, a much stronger indicator that a word might be a good keyphrase for domain-independent detection of offensive speech. Indeed, the list of the 10 highest-ranked nouns barely need post-editing, they are *Vasall* 'vassal', *Invasoren* 'invaders', *Abschaum* 'scum', *Heuchler* 'howler', *Dumm* 'dumb', *Kreatur* 'creature', *Ficker* 'fucker', *Titten* 'boobs', *Scheisse* 'shit', *Volksverräterin* 'traitor of the people'. We note that emoji characters are also handled by stanza as individual words and some of them also appear in the final keyword lists, such as these characters: 🤮🤬🖕💩 . The extraction and ranking of patterns is implemented in the `ml` module of the `tuw-nlp` library, the manually curated rule lists and code to apply them are part of the `tuw-inf-germeval2021` repository.

In an independent effort we also used the 2021 training dataset for all three subtasks to observe simple patterns that can be used as high-precision predictors of each category. Two of the patterns we identified were introduced in the final rule-based system: for the toxicity detection task we categorize a comment as toxic if it contains at least two words with at least four characters each written in ALL-CAPS. This rule on its own achieved 91% precision and 4% recall on our validation set. For Subtask 3, if a comment contains an URL and this URL is not the only content of the comment, we categorize it as fact-claiming. This rule achieved 93% precision and 10% recall on the validation set.

## 5 Ensemble

Our three submissions for the shared task are combinations of the systems described in Sections 3 and 4. Our first submission contains the decisions of the BERT-based classifiers for each subtask, using the number of iterations determined as yielding the highest F-score (2nd for Subtask 1, 1st for Subtask 2, and 5th for Subtask 3). Submission 2 is the union of Submission 1 and our rule-based systems, i.e. for each subtask we label comments as toxic/fact-claiming if either the BERT-based model or our rule-based system would classify it as such (we did not use any rules for subtask 2). Finally, Submission 3 is our attempt at a system with higher precision at the cost of recall, here we use our rules together with BERT models from the iterations yielding the highest precision (8th for Subtask 1, 1st for Subtask 2, and 1st for Subtask 3). We note that this is different from training a machine learning model for high precision, which could be achieved by e.g. a weighted loss function. In case of Subtask 2, both precision and F-score were optimal after the same number of training epochs. Since we did not use any rules for detecting engaging comments, our output for this subtask was identical in all three submissions.

## 6 Quantitative results

Quantitative evaluation of our methods is performed based on the test set provided by the organizers. We follow the official evaluation methodology and calculate precision, recall, and F-score for both classes in each subtask and also the macro-average across classes for each figure. Results on toxicity detection (Subtask 1) are presented in Table 1. Our rule-based system did not achieve higher precision than



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

the BERT-based system, but it selected a somewhat different set of comments and increased the recall and F-score of the ensemble system in Submission 2. We shall take a closer look into the contributions of the rule-based system as part of our qualitative analysis in Section 7. The BERT model chosen for high precision performed worse, possibly because of overfitting on the training dataset (it was trained for 8 epochs as opposed to the 2 epochs of the high F model).

For the task of detecting engaging comments (Subtask 2) we did not develop any rule-based system and the same BERT model was determined to be optimal for both precision and F-score, therefore we used the output from the same BERT model in all our submissions. Here we omit results for this subtask due to lack of space. Results on detecting fact-claiming comments (Subtask 3) are presented in Table 2. Although the rule-based system achieves high precision, the comments it identifies as fact-claiming (based on the single rule regarding URLs, see Section 4) form a subset of the comments identified as such by the BERT model, hence our labels for Submissions 1 and 2 are identical. The BERT model chosen for high precision indeed makes very few false positive decisions and can be slightly improved in terms of both precision and recall by adding labels from the rule-based system (Submission 3).

## 7 Qualitative analysis

The main focus of our rule-based experiments was the toxicity detection. We performed a detailed qualitative analysis on a sample of the test dataset on this subtask. Based on the labels assigned by the BERT model of Submission 1 and the ground truth labels we extracted a sample of 40 comments, 10 from each of the four categories true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Our goal with this setup is to go beyond error analysis, which would only focus on false positives and false negatives. Given the subjectivity of the task and of the possible discrepancies between the information available to annotators and to our models (see Section 2), we wished to inspect a sample that is balanced across both predicted and ground truth labels. Since the BERT model of Submission 3 performed poorly, probably due to overfitting, for the purposes of this analysis we shall focus on the output of the BERT model of Submission 1 (trained for 2 epochs, for maximum F-score) and our rule-based system. The number of comments in each category for both of these systems and their combination is presented in Table 3. Figure 1 shows the comments in our samples of TP, FP, TN, and FN predictions of the BERT model, respectively. An asterisk (*) marks an agreement with the rule-based system, e.g. TP1 and TP2 were classified as toxic by both models and TP3-10 were classified as toxic by BERT but not by the rule-based system, and all ten have been labeled as toxic by the annotators (hence *true* positive).

The sample of true positives, as expected, contains many comments that are clearly identifiable as such based on surface patterns such as the word *dumm* 'stupid' (TP1, TP10) or the emojis 💩 and 👎 (TP2). False negatives (FN) would be expected to exhibit the opposite pattern, these are comments that humans agreed are toxic but models failed to detect them as such. Indeed this group contains several examples where a deep understanding of the comment is necessary to account for its toxicity, demonstrating the complexity of the task. For example, to understand that comment FN9 *Der Deutsche war schon immer naiv…* 'The Germans have always been naive' is in some way uncivil, one must know that some types of statements about some types of groups are not acceptable and at the same time be able to identify *Germans* and *naive* as concepts belonging to these 'types'. Indeed, if a human expert were to build a complex rule-based model for the toxicity detection task, it may very well contain patterns such as *PROTECTED_GROUP + NEGATIVE_PREDICATE* and lexica for what words and phrases are to be considered as belonging to each of these categories. Other examples require knowledge of idioms, such as the comment FN1 *Geht's noch?* which as a phrase can be translated as 'Are you crazy?'[6]. Perhaps the most puzzling examples of false negatives are those that were probably interpreted as sarcastic by annotators, such as FN2 *Good Luck Mr. President Trump* 👍👋❤️ *Make America Great Again. Ich würde ihn wählen.*

[6] https://de.wiktionary.org/wiki/geht%E2%80%99s_noch



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

| | |
|---|---|
| TP1* | Meine Güte redet die Dame Quatsch! Frauen sind also Feiglinge und Männer zu dumm . Aha |
| TP2* | Putin ist sehr nerwös 😎💨💩😂👎 |
| TP3 | @USER Ja. Er ist blödsinnig. |
| TP4 | Schon wieder ein klugscheiss!! |
| TP5 | DANKE Carla super weiter so ❤️❤️❤️ |
| TP6 | Trump 👍🏾👍🏾👍🏾👍🏾🤣🤣 Biden 🇨🇳🇨🇳🇨🇳🇨🇳🇮🇷🇮🇷🇮🇷 |
| TP7 | Was war jetzt so schlimm in den letzten vier Jahren? Unter Trump bleibt die Welt friedlich. Gewinnt Biden, wirds Krieg geben. Guten Tipp an Zamperoni, verlass deine Frau! |
| TP8 | 😂🤣👨‍👨‍👨‍ |
| TP9 | Covid Karls Paniktour 😂😂😂😂 |
| TP10 | @USER seit ihr so naiv, oder tut ihr nur so dumm?????? 🎩🌹 |
| FP1* | @USER es gibt halt leider in jeder Altersgruppe asoziale Mitmenschen. |
| FP2* | @USER Gott sei Dank gibt es eine Meinungsfreiheit und verschiedene Auffassungen. Und ganz nebenbei gefragt, was hat diese Dame denn für Vorschläge gemacht? |
| FP3* | Frau ”Blaerbock” in ihrem Element. Fehlt nur noch eine Horde ”Kobolde”! |
| FP4 | @USER stimmt schon, gegen den würde Reagen wie Bernie Sanders wirken. |
| FP5 | pandemie? 🤣 97% der deutschen sind weder infiziert, noch krank! |
| FP6 | Schnell viel Blödsinn reden… |
| FP7 | Sitzt da die Schwester von Didi Hamann(Augenzucken)? |
| FP8 | …keine ahnung…woll?!? 😉 |
| FP9 | Dünnes Eis Frau Prof. |
| FP10 | Tja, mit Ideologie wird es kalt und dunkel hier! 🤦‍♀️ |
| TN1* | @USER woher wissen Sie das? Glauben heißt nicht Wissen 🤓 |
| TN2* | Sie wissen schon dass dies in der Konsequenz vielen Menschen die sich schutzlos infizieren das Leben kosten wird ? |
| TN3* | @USER paint it red |
| TN4* | @USER weil er es kann. 🤷‍♂️ |
| TN5* | @USER ne 🙈 |
| TN6* | @USER das ist leider auch wahr |
| TN7* | Wer’s glaubt 👋👍 |
| TN8* | @USER sehr guter Kommentar 👍 |
| TN9* | 👍 Trump 👍 |
| TN10* | Weiterhin gute Besserung! |
| FN1* | Geht’s noch? |
| FN2* | Good Luck Mr. President Trump 👍👋❤️ Make America Great Again. Ich würde ihn wählen. |
| FN3* | Tourett? |
| FN4* | Politiker mit Verstand ?…und das in Zusammenhang mit dem Namen Trump ? |
| FN5* | Und wenn ich mir die Frage stelle, ob ich einen Arbeitsplatz habe und die Miete bezahlen kann, dann wähle ich Biden? 😂😂😂 |
| FN6* | Trump wird gewinnen und das ist gut so. Auf die Gesichter der Hetzmedien bin ich gespannt |
| FN7* | @USER immer politisch Korrekt verstehst Du 😉😉 |
| FN8* | @USER viel Meinung für wenig Ahnung! |
| FN9* | @USER Der Deutsche war schon immer naiv, und sie sind es auch. |
| FN10* | Mimimimi |

Figure 1: Sample of 10 comments each from true positive (TP), false positive (FP), true negative (TN), and false negative (FN) labels predicted by the BERT model on the test set for Subtask 1. An asterisk (*) marks comments where the rule-based system made the same prediction as the BERT model

| | Other | | | Toxic | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| Rules | 64.0 | **98.3** | 77.5 | **67.7** | 6.0 | 11.0 | 65.9 | 52.2 | 58.2 |
| BERT (S1) | 72.4 | 87.9 | 79.4 | **67.7** | 43.1 | 52.7 | 70.1 | 65.5 | 67.7 |
| BERT + Rules (S2) | **73.2** | 87.0 | **79.5** | 67.6 | **46.0** | **54.8** | **70.4** | **66.5** | **68.4** |
| BERT-high-prec | 71.9 | 86.4 | 78.5 | 64.9 | 42.9 | 51.6 | 68.4 | 64.6 | 66.5 |
| BERT-high-prec + Rules (S3) | 72.9 | 85.9 | 78.8 | 65.6 | 45.7 | 53.9 | 69.2 | 65.8 | 67.5 |

Table 1: Results on Subtask 1

| | Other | | | Fact-Claiming | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| Rules | 68.0 | 99.7 | 80.8 | 90.0 | 5.7 | 10.8 | 79.0 | 52.7 | 63.2 |
| BERT (S1) | **83.8** | 74.9 | 79.1 | 58.5 | **71.0** | 64.2 | 71.2 | **73.0** | **72.1** |
| BERT + Rules (S2) | **83.8** | 74.9 | 79.1 | 58.5 | **71.0** | 64.2 | 71.2 | **73.0** | **72.1** |
| BERT-high-prec | 71.0 | **99.4** | **82.8** | 93.5 | 18.5 | 30.9 | 82.3 | 58.9 | 68.7 |
| BERT-high-prec + Rules (S3) | 71.1 | **99.4** | **82.9** | **93.7** | 18.8 | 31.3 | **82.4** | 59.1 | 68.8 |

Table 2: Results on Subtask 3

| | **TP** | **FP** | **TN** | **FN** |
|---|---|---|---|---|
| BERT | 151 | 72 | 522 | 199 |
| Rules | 21 | 10 | 584 | 329 |
| BERT + Rules | 161 | 77 | 517 | 189 |

Table 3: Number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) labels from each of our systems on the test set for Subtask 1

'I would vote for him'. A similar example is TP5: *DANKE Carla super weiter so* ❤️❤️❤️ 'Thanks Carla super keep it up', which may have been detected by the BERT model because of the capitalized word, but to account for the positive label in the ground truth we can only speculate once again that annotators have interpreted it as sarcastic.

Turning to comments that were not labeled as toxic by annotators, in the sample of true negatives (TN) we did not find any controversial examples. The false positives (FP), on the other hand, once again provide examples of the inherent difficulty and subjectivity of the task. Consider e.g. FP6 *Schnell viel Blödsinn reden…* 'Quickly talk a lot of nonsense…' and FP10 *Tja, mit Ideologie wird es kalt und dunkel hier!* 🤦‍♀️ 'Well, with ideology it gets cold and dark here!'. We believe that it would be necessary to also consider the post fragments that annotators had access to but were not included in the dataset (see Section 2) to determine how such comments could have been unanimously labeled as non-toxic. Several FP examples, however, are clearly not toxic, and in case of BERT one can only speculate as to why they were falsely classified as such. The three comments that were also false positives of the rule-based system (FP1-3) were misclassified because of the presence of the words *asozial* 'asocial', *Meinungsfreiheit* 'freedom of opinion', and *Horde* 'horde', illustrating the limitations of purely keyword-based methods.

The analysis in this section was intended to provide examples of the types of challenges a model of toxicity must concern itself with. While it is limited to a small sample from a single dataset, we believe it illustrates a range of problems that are typical for this task. In particular, false negative predictions are responsible for more than 70% of errors made by both of our top-performing systems, and our analysis suggests that identifying most of these would require more complex rules for modeling specific types of toxicity and the ability to detect sarcasm.

## 8 Conclusion

We described simple methods for the semi-automatic construction of rule-based systems for detecting toxicity in social media, and used them to improve the performance of a BERT-based classifier on the dataset of the 2021 GermEval shared task. A manual error analysis was provided to illustrate the most challenging aspects of the task.

## Acknowledgments

Research conducted in collaboration with Botium GmbH.

## References


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Tim Graf and Luca Salini. 2019. bertzh at germeval 2019: Fine-grained classification of german offensive language using fine-tuned bert. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 434–437, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Andrei Paraschiv and Dumitru-Clementin Cercel. 2019. Upb at germeval-2019 task 2: Bert-based offensive language classification of german tweets. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 398–404, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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.

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 *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 352–363, München, Germany. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

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.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*, pages 1–10, Vienna, Austria. Austrian Academy of Sciences.

Lena Wilms. 2021. personal communication.