Offensive Text Detection Across Languages and Datasets Using Rule-based and Hybrid Methods

Anonymous ACL submission

Abstract

We investigate the potential of rule-based systems for the task of offensive text detection in English and German, and demonstrate their effectiveness in low-resource settings, as an alternative or addition to transfer learning across tasks and languages. Task definitions and annotation guidelines used by existing datasets show great variety, hence state-of-the-art machine learning models do not transfer well across datasets or languages. Furthermore, such systems lack explainability and pose a critical risk of unintended bias. We present simple rule systems based on semantic graphs for classifying offensive text in two languages and provide both quantiative and qualitative comparison of their performance with deep learning models on 5 datasets across multiple languages and shared tasks.

1 Introduction

The task of offensive text detection, especially as applied to social media has seen a rise of interest in recent years, with many overlapping definitions of categories such as toxicity, hate speech, profanity etc. Datasets are constructed using different sets of class definitons corresponding to different annotation instructions, and machine learning models that learn patterns of one dataset may perform poorly on another. Modern deep learning models also offer little or no explainability of their decisions, and their potential for unintended bias reduces their applicability in real-world scenarios such as automatic content moderation. In this paper we present a rule-based approach, a semi-automatic method for constructing patterns over Abstract Meaning Representations (AMR graphs) built from input text, and evaluate its potential as an alternative to machine learning for offensive text detection across two languages (English and German) and 5 datasets from 2 shared tasks. Our quantitative analysis compares the rule-based method to both monolingual and multilingual deep learning models trained on data from each language and shared task, demonstrating its potential in low-resource settings as an alternative or addition to transfer learning. Our qualitative analysis examines the decisions made by each system on samples of 100-100 texts from both languages and provides a subjective categorization of their errors to demonstrate the sensitivity of quantitative evaluation to the characteristics of individual datasets and their potentially controversial annotations. The rest of this paper is organized as follows. An overview of related work and the most important shared tasks and datasets is given in Section 2, the datasets used in our experiments are described in Section 3. Our method for constructing AMR-based rule systems is presented in Section 4 and our experiments are described in Section 5. Quantitative evaluation is presented and discussed in Section 6, the qualitative analysis on samples from two datasets is provided in Section 7. All software for experiments as well as the rule-based systems presented is available as open-source software under an MIT license from `https://anonymous.4open.science/r/offensive_text`.

2 Related Work

Datasets As pointed out already in a 2017 survey (Schmidt and Wiegand, 2017), the definition of offensive text varies greatly across datasets, which makes the portability of deep learning models for offensive text detection a hard problem. Annual shared tasks on hate speech detection and related tasks may use similar definitions year after year, but there is great variation when moving from one shared

task to another and models that achieve high quantitative results on their targeted test set don't generalize well (see (Yin and Zubiaga, 2021) for a recent survey). In this paper we shall experiment on yearly datasets from two tasks that both use the same labeling scheme for offensive text, HASOC (Mandl et al., 2021) and GermEval (Risch et al., 2021). Both challenges define a binary classification of social media texts (Tweets or Facebook comments) into the *offensive* and *non-offensive* classes, and a fine-grained classification of the offensive category into the subclasses *abusive*, *insulting*, and *profane*. A detailed description of these tasks and datasets will be given in Section 3. The OLID and SOLID datasets of SemEval 2019 (Zampieri et al., 2019) and 2020 (Zampieri et al., 2020) use task definitions similar to GermEval. Other widely used datasets with a narrower scope include the data provided by the TRAC (Kumar et al., 2018, 2020b) and HatEval (Basile et al., 2019) shared tasks. TRAC contains English, Hindi and Bangla data from Twitter and Facebook and annotation focuses on the categories aggression and misogyny, the HatEval task is concerned with hate speech directed at immigrants or women in English and Spanish Twitter data.

Approaches Most systems for offensive text detection rely on distributional text representations, including both static (Chiril et al., 2019) and contextual embeddings (Kumar et al., 2020a; Ranasinghe and Zampieri, 2020). As in many popular text classification tasks, the most widely used neural language models are based on the Transformer architecture (Vaswani et al., 2017), and in particular BERT-based models (Devlin et al., 2019) are the basis of the state of the art machine learning systems for most datasets, including the best-performing systems on GermEval2021 (Bornheim et al., 2021), GermEval2019 (Paraschiv and Cercel, 2019), HASOC 2020 English (Ghanghor et al., 2021) and HASOC 2020 German (Kumar et al., 2020a; Dowlagar and Mamidi, 2021) Top systems enhance quantitative performance by optimizing metaparameters such as maximum sentence length or number of training epochs (Kumari and Singh, 2020; Liu et al., 2019), by training on joint subtask labels (Mishra and Mishra, 2019) or utilizing multiple Transformer based models to counteract the small dataset sizes (Bornheim et al., 2021), by pre-training on additional hate speech corpora (Caselli et al., 2020), training jointly on different corpora (Kumar et al., 2020a), or by using adversarial learning (Tran et al., 2020). Further deep learning methods used in offensive text detection include LSTMs (Wang et al., 2019; Mishra et al., 2020), CNNs (Gambäck and Sikdar, 2017; Park and Fung, 2017), or both (Badjatiya et al., 2017), sentence embeddings (Indurthi et al., 2019), and ensembles of multiple machine learning models (Badjatiya et al., 2017; Nikolov and Radivchev, 2019).

Explainability and rule learning The interpretability of NLP models and the explainability of their decisions is subject of growing interest, also as part of the broader research area of explainable artificial intelligence (xAI). Deep learning models are considered black boxes in most applications and efforts to interpret them are generally limited to feature weight visualizations with limited validity (see e.g. (Serrano and Smith, 2019), (Wiegreffe and Pinter, 2019), and (Jain and Wallace, 2019) for the controversy about using attention weights as explanation). Yet even the more mature methods for interpreting neural networks (e.g. LIME (Ribeiro et al., 2016)) do not offer the kind of transparency of ML models that would allow developers to customize their functionality the way a domain expert can update a traditional rule system. In this work we experiment with a rudimentary method for semi-automatic, human-in-the-loop (HITL) learning of simple rule systems over semantic graphs. Recent approaches to automatic learning of rule systems for NLP tasks range from the learning of first order logic formulae over semantic representations using neural networks (Sen et al., 2020) and integer programming (Dash et al., 2018) to the training of probabilistic grammars over semantic graphs (Donatelli et al., 2019). Human-in-the-loop (HITL) approaches involve generating rule candidates to be reviewed by experts, e.g. by extracting textual patterns (Lertvittayakumjorn et al., 2021) or semantic struc-

tures (Sen et al., 2019) Rule-based approaches are also often combined with ML methods, e.g. by incorporating lexical features into DL architectures (Koufakou et al., 2020; Pamungkas and Patti, 2019) or voting between rule-based and ML systems (Razavi et al., 2010; Gémes and Recski, 2021).

3 Data

In this section we introduce datasets from the GermEval and HASOC shared tasks, which are the basis of all our quantitative experiments in Section 5 and our qualitative analysis in Section 7. Our choice of two recent tasks that use identical labeling schemes and both include data for both English and German, two languages in which we are able to perform qualitative analysis (see Section 7), allows us to investigate the ability of both ML and rule-based models to transfer between tasks as well as languages.

GermEval The GermEval shared task was organized in 2018 (Wiegand et al., 2018), 2019 (Struß et al., 2019), and 2021 (Risch et al., 2021). German Twitter posts were annotated for the 2018 and 2019 challenges, the 2021 task used comments from a news-related Facebook group. Besides the detection of offensive texts, the 2021 Germeval offered annotations for two additional tasks, the identification of engaging and fact-claiming comments, but these datasets are not part of our current experiments. The 2018 and 2019 Twitter datasets consist of posts from 100 user timelines and is limited to tweets in German that are not retweets, do not contain URLs, and contain at least 5 alphabetic tokens. The dataset is not a random sample of posts meeting these criteria, users were heuristically selected to ensure a high ratio of offensive tweets (further details on this selection were not given), then the dataset was debiased using additional tweets with non-offensive words that were observed to be overrepresented in offensive posts, such as *Merkel* or *Flüchtlinge* 'refugees'. The 2021 edition of Germeval featured a collection of comments from the Facebook page of a German political talk show. The 2021 training data was collected between January and June of 2019, while the test set is from between September and December of 2020. The dataset has been anonymized to comply with Facebook's guidelines for publishing data. The datasets from 2018 and 2019 categorize the offensive texts further into three categories, *profanity, insult*, and *abuse* and defines offensive text as the union of these categories, this is identical to the definition used at HASOC. The 2021 dataset does not contain such fine-grained labels and defines offensive texts as the union of *screaming, vulgar language, insults, sarcasm, discrimination, discrediting*, and *accusation of lying*.

HASOC The Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) shared task was inspired by GermEval and OffensEval and was organized in 2019 (Mandl et al., 2019), 2020 (Mandl et al., 2020), and 2021 (Mandl et al., 2021). The dataset from 2019 contained tweets and Facebook comments in English, Hindi and German. Offensive posts were selected based on keywords and hashtags, and debiased similarly to the process described by GermEval organizers. From 2020 datasets were selected by training a Support Vector Machine classifier (SVM) on a collection of hate speech datasets and using this classifier to select the tweets to be annotated for the dataset. Following the definition of the 2019 and 2020 GermEval challenges, each HASOC task distinguishes between three types of offensive text, those displaying profanity (`PRFN`), offense (`OFFN`), or hate (`HATE`). The binary classification of offensive texts considers the union of these three categories, and both our quantitative experiments in Section 5 and our qualitative analysis in Section 7 are concerned with this task only.

4 Method

In our quantitative experiments as well as in our error analysis we compare the performance of standard deep learning models with rule-based systems that define sets of patterns over AMR graphs built from the texts of posts to be classified. For the DL models we use standard architectures without modification, technical details will be described along with the experimental setup in Section 5. In this section we describe our rule-based approach.

Abstract Meaning Representations (AMR,

(Banarescu et al., 2013)) are directed graphs of concepts representing the semantics of a sentence. We construct AMR graphs for each of our datasets using neural text-to-AMR parsers. For English we use a pre-trained Transformer-based AMR parser (Raffel et al., 2020) and the `amrlib`[1] library, for German we construct AMRs from text using a multilingual, transition-based (Damonte and Cohen, 2018) system via the `amr-eager-multilingual`[2] library. Our rule system for each task is a list of patterns over AMR graphs, and applying such a rule system to a piece of text means labeling it as offensive iff at least one pattern in the list matches the corresponding AMR graph. Individual patterns are graphs whose edge and node labels may be strings or regular expressions (regexes) defining sets of possible labels, and a graph pattern with regexes for labels defines the set of all graphs whose corresponding node and edge labels are matched by those regexes. Patterns can also be negated and a conjunction of patterns used as a single rule, a complete rule system can therefore be considered a single boolean statement in disjunctive normal form (DNF) of boolean predicates corresponding to graph patterns, in this regard our method is similar to the approach of (Dash et al., 2018) and (Sen et al., 2020) (see Section 2).

To construct rule systems efficiently, we implement a form of human-in-the-loop (HITL) learning. For each training dataset we consider all AMR graphs and generate a list of frequently occurring subgraphs with at most 2 edges, then rank them based on their importance for the classification task. For this we use subgraphs as features to train a decision tree on the dataset using the `sklearn` library and then rank these features based on their Gini coefficient. The maximum size of subgraphs is a free parameter of the system but must be kept low to limit the search space. We thus obtain a ranked list of relevant graph patterns that we can use to construct our rule system manually, by building a list of rules, each of which is either a single pattern or the conjunction of multiple patterns, any of which may be negated. We shall describe the individual rule systems built for our experiments in Section 5.

[1] `https://amrlib.readthedocs.io/en/latest/`

[2] `https://github.com/mdtux89/amr-eager-multilingual`

5 Experiments

Quantitative evaluation is performed using 5 datasets. For English we train models using the three datasets from the 2019-2021 editions of the HASOC shared task, for German we use the 2021 GermEval dataset (the training portion of which is from earlier editions of GermEval) and the 2020 HASOC corpus (see Section 3 for details on each dataset). We train standard BERT-based classifiers on each dataset and compare them with rule systems we built manually. We investigate the ability of models to transfer between tasks by evaluating each of them on the test sets of all other datasets as well. We also attempt transfer learning across languages (English to German and German to English), by training models using multilingual BERT on datasets from one language and evaluating them on the other language. Finally, we also measure the contribution of our rule-based system to DL models by evaluating the union of their predicted positive labels, i.e. by considering the strategy of classifying a text as offensive iff at least one of multiple models would classify it as such. In this section we provide details of our deep learning experiments, followed by an overview of our rule systems built from each dataset using the method in Section 4. Results and discussion follow in Section 6.

Deep learning models For training BERT-based models we preprocess text data by replacing emoticons with their textual representation using the *emoji*[3] Python library, then removing hashtag symbols and substituting currencies and urls with special tags using the regex-based library *clean-text*[4]. Finally, we use our own regular expressions for masking usernames, media tags, and moderators, by replacing each with the *[USER]* tag. For both languages we fine-tune a language specific pretrained BERT model (`bert-base-german-cased`[5] for German and `bert-base-uncased`[6] for

[3] `https://pypi.org/project/emoji/`

[4] `https://pypi.org/project/clean-text/`

[5] `https://deepset.ai/german-bert`

[6] `https://huggingface.co/bert-base-uncased`

English) as well as the multilingual model (`bert-base-multilingual-cased`[7]). On each dataset we then train one model with the language-specific BERT and one with multilingual BERT. Each of the 6 datasets consists of a train and test portion. For selecting training metaparameters we further divide the train portions of each dataset into into train and validation sets, using a 3:1 ratio, then for the final experiments we train our models using the full training datasets and evaluate them on the test sets. For each dataset we train a neural network with a single linear classification head on top of BERT. Metaparameters are set based on performance on the validation set. We use Adam optimizer with a weight decay value of 10^{-5} and initial learning rate of 10^{-5}. We use the balanced weighted loss function of *sklearn*[8], to compensate for unbalanced labels, as suggested by (King and Zeng, 2001). We set batch size to 8 and train each model for 10 epochs, then determine the optimal number of iterations based on their F-score on the validation set[9].

Rule based system For building and applying our AMR-based rule systems we parse all text with language-specific text-to-AMR parsers (see Section 4 for details). The only preprocessing step we apply is the replacement of emoticons, as described in the previous paragraph. We build rule systems based on each of the 5 training datasets (HASOC 2019-2021 for English, GermEval 2021 and HASOC 2020 for German). Rule systems were built semi-automatically by the authors, based only on the training portions of each dataset, test sets were excluded from the process entirely and even validation sets were only used for quantitative evaluation, but not for HITL learning or manual analysis. All rule lists are presented in their entirety in Appendix A, here we provide an overview only. In each of the 5 rule systems the rules with the highest yield are those that consist of a single node, i.e. that refer to the presence of a single word in the text. The majority of these words are in themselves profane and/or insulting. In English rule systems top keywords include *asshole, stupid, bitch, shit, fuck* as well as *useless* and *disgrace* (see Appendix A for full lists). In German rule sets the top words that trigger the offensive label in themselves also include *ficken* 'fuck', *porno*, *hurensohn* 'son of a bitch', *arsch* 'ass' and *scheiße* 'shit'. Rules with multiple nodes typically serve to separate offensive and non-offensive occurrences of a word. For example, the word *shame* is present in over 200 offensive posts of the English HASOC 2021 dataset, but as a keyword rule it would also yield 43 false positives. Using a pattern over AMR graphs we can filter occurrences of the word by the object (*ARG1*) of *shame* and construct the rule *shame* $\xrightarrow{ARG1}$ *(media/person/publication/they/you/party/have/government)*, which yields only 8 false positives for 103 true positives. Another example of patterns over multiple nodes are rules covering negation. For example, in the rule system based on the GermEval 2021 training set, the rule *normal* $\xrightarrow{polarity}$ – matches all posts where the word *normal* is negated, such as in the sentence *Das ist doch nicht mehr normal!* 'That's just not normal anymore!'. The complete rule lists built from each of the 5 datasets are presented in Appendix A.

6 Results

The shared tasks we focus on each evaluate classifiers by measuring precision, recall, and F1-score on both the offensive and non-offensive class, and systems are ranked based on the macro-average F-score, which is not the average of two F1-scores but the harmonic mean of the macro-average precision and recall scores. HASOC organizers argue that using macro-average F1-score counteracts class imbalance (Mandl et al., 2019). We follow this practice in our evaluation, especially since many of the top participating systems do not publish scores for individual classes, thus we can only compare our models to theirs using the macro-average F-score. Our main results on the test portions of each of the 5 corpora is presented in Table 1. On each dataset we

[7] https://huggingface.co/bert-base-multilingual-cased

[8] https://scikit-learn.org/

[9] the optimal number of epochs for each model were the following: EN: 2, EN-multi: 2, DE: 1, DE-multi: 3, DE-HASOC: 2, DE-HASOC-multi: 8, GermEval: 6, GermEval-multi: 4

evaluate DL models trained on data from the same task, on data from the other task of the same language, on all data in the language, or on all data from the other language (using multilingual BERT). Additionally we evaluate our dataset-specific rule systems and the pairwise unions of various systems. Where available, we also present the (macro-average) scores of the top-performing system for each dataset, as well as those of the original winner of each shared task (if the two differ). Additional quantitative results are presented in Appendix B.

As expected, the best results are achieved by DL models trained on data from the same task, with or without additional data from the same language. These models are typically within a few percentage points of the best models, and are not improved significantly with the addition of the rule system. It is also expected that rule systems achieve the highest precision values on each dataset, this is by design and at the expense of recall. The effect of rules as an enhancement is considerable in the case of the transfer learning scenarios, both between task and between languages. Since rules are generally high-precision, most models' performance is improved by considering their union with the task-specific rule system (taking the union of two or more binary classifiers means classifying a text as offensive iff at least one of the models classifies it as such). This effect can be observed on both German and English datasets. On the German HASOC dataset, where the EN-multi model trained on English data only is in itself more than 20 points below the F-score on the offensive class achieved by the model trained on the training data corresponding to the test set (DE-HASOC), but adding labels predicted by the rule-based system closes almost half of this gap, raising F-score from 52.9 to 61.9. On the 2019 English HASOC dataset the effect is similar, rules close about half of the performance gap between German and English models. This effect shows the potential of simple rule systems in low-resource scenarios where training data is only available for other languages and/or for other tasks/genres. On some datasets, our rule systems work well as standalone solutions as well. In case of the 2020 English dataset our rules achieve 83.7 F-score on the offensive class, compared to 90.3 of the best DL system. We believe that in real-world applications, e.g. automatic content moderation, such a system may be preferred despite its lower performance, due to its transparency and the fact that its precision is above 95% compared to the 90% of the top black-box system.

7 Error Analysis

In this final section we perform manual error analysis on samples of 100 posts each of the 2021 datasets for each language (GermEval for German and HASOC for English). Samples were selected randomly and classified by each of the models described and evaluated in previous sections. Here we provide an overview of errors made by each model and cite selected examples. Errors made by our models are grouped into what we consider to be typical error classes, but we note that such a categorization is subjective and is made solely for the purpose of discussion and presentation of the results of our manual analysis. The examples we refer to in our discussion below are presented in Table 2, a full list of errors made by each of the systems as well as quantitative evaluation of each classifier on the two samples is available in Appendix E.

The largest error class consists of false negative predictions that are clearly offensive and some models failed to detect them as such. These include e.g. the profanity in FNen14†‡ or the insult in FNde1*†‡. Another major group consists of posts on controversial/sensitive topics whose status as offensive/non-offensive is influenced by both form and content and is also probably controversial. False positive predictions in this group include texts that express strong negative opinions in a relatively civil way (FPde2*, FPde4*), while false negatives are those that may have been annotated as offensive because of their tone (FNen1*†‡, FNen2*†‡). Ground truth annotations are inconsistent about whether the presence of profanity alone warrants the offensive label. The posts FPen1*‡ and FPen2*†, which have been predicted as offensive by several of our models and contain words such as *fuck* and *bitch*, are annotated as non-offensive.

Test	System	Offensive			Other			Macro avg		
		P	R	F	P	R	F	P	R	F
DE GermEval2021	Rules	65.4	9.7	16.9	64.6	**97.0**	77.5	65.0	53.3	58.6
	DE-All	**72.9**	35.4	47.7	70.8	92.3	**80.1**	71.9	63.8	67.6
	DE-GermEval	56.7	48.6	52.3	72.0	78.1	75.0	64.4	63.3	63.8
	DE-HASOC	69.6	11.1	19.2	65.0	97.1	77.9	67.3	54.1	60.0
	EN-multi	53.4	20.0	29.1	65.6	89.7	75.8	59.5	54.9	57.1
	DE-All ∪ Rules	69.8	40.3	51.1	71.8	89.7	79.8	70.8	65.0	67.8
	EN-multi ∪ Rules	54.9	27.4	36.6	67.0	86.7	75.6	60.9	57.1	58.9
	DE-All ∪ EN-multi	62.3	44.9	52.2	72.1	84.0	77.6	67.2	64.4	65.8
	DE-All ∪ EN-multi ∪ Rules	60.9	**48.9**	**54.2**	**73.0**	81.5	77.0	66.9	65.2	66.0
	FHAC	-	-	-	-	-	-	**73.1**	**70.4**	**71.8**
DE HASOC2020	Rules	**92.4**	28.3	43.4	77.0	**99.0**	86.6	**84.7**	63.7	72.7
	DE-All	55.4	93.0	69.4	96.0	69.1	80.3	75.7	81.0	78.3
	DE-GermEval	47.7	90.7	62.5	93.9	59.0	72.5	70.8	74.8	72.8
	DE-HASOC	66.6	81.7	**73.4**	91.7	83.1	**87.2**	79.1	**82.4**	**80.7**
	EN-multi	57.4	49.0	52.9	80.2	85.0	82.5	68.8	67.0	67.9
	DE-All ∪ Rules	55.4	93.3	69.6	96.2	69.1	80.4	75.8	81.2	78.4
	EN-multi ∪ Rules	62.1	61.7	61.9	84.2	84.5	84.3	73.2	73.1	73.1
	DE-All ∪ EN-multi	51.1	94.7	66.4	96.6	62.6	76.0	73.8	78.6	76.2
	DE-All ∪ EN-multi ∪ Rules	51.2	**95.0**	66.5	**96.8**	62.6	76.0	74.0	78.8	76.3
	ComMA	-	-	-	-	-	-	-	-	52.4
	HASOCOne	-	-	-	-	-	-	-	-	77.9
EN HASOC2021	Rules	**87.2**	45.1	59.5	49.5	89.0	63.7	68.4	67.1	67.7
	EN	80.3	95.2	**87.2**	88.7	61.5	**72.6**	**84.5**	**78.4**	**81.3**
	DE-All-multi	82.7	23.9	37.1	42.2	**91.7**	57.8	62.4	57.8	60.0
	EN ∪ Rules	79.8	95.6	87.0	**89.2**	60.0	71.8	**84.5**	77.8	81.0
	DE-All-multi ∪ Rules	84.1	53.9	65.7	52.2	83.2	64.2	68.2	68.6	68.4
	EN ∪ DE-All-multi	79.3	95.5	86.6	88.8	58.8	70.7	84.0	77.1	80.4
	EN ∪ DE-All-multi ∪ Rules	78.8	**95.7**	86.4	89.1	57.3	69.8	83.9	76.5	80.1
EN HASOC2020	Rules	**95.3**	74.6	83.7	78.6	**96.2**	86.5	86.9	85.4	86.2
	EN	90.2	90.5	**90.3**	90.2	89.9	**90.1**	**90.2**	**90.2**	90.2
	DE-All-multi	79.3	20.9	33.1	53.7	94.4	68.5	66.5	57.7	61.8
	EN ∪ Rules	89.6	91.0	**90.3**	90.6	89.2	89.9	90.1	90.1	90.1
	DE-All-multi ∪ Rules	89.8	78.7	83.9	80.6	90.8	85.4	85.2	84.8	85.0
	EN ∪ DE-All-multi	86.6	91.9	89.2	91.2	85.4	88.2	88.9	88.6	88.8
	EN ∪ DE-All-multi ∪ Rules	86.0	**92.3**	89.1	**91.5**	84.6	87.9	88.7	88.5	88.6
	IIIT_DWD	-	-	-	-	-	-	-	-	51.5
	IIITK	-	-	-	-	-	-	-	-	**93**
EN HASOC2019	Rules	**73.2**	35.1	47.4	81.6	**95.7**	**88.1**	**77.4**	65.4	70.9
	EN	59.6	76.7	**67.1**	91.4	82.7	86.8	75.5	**79.7**	77.5
	DE-All-multi	53.1	47.9	50.4	83.2	85.9	84.5	68.1	66.9	67.5
	EN ∪ Rules	57.5	77.4	66.0	91.5	80.9	85.9	74.5	79.2	76.8
	DE-All-multi ∪ Rules	55.0	63.5	58.9	87.2	82.7	84.9	71.1	73.1	72.1
	EN ∪ DE-All-multi	51.5	82.6	63.5	**92.8**	74.1	82.4	72.1	78.4	75.1
	EN ∪ DE-All-multi ∪ Rules	50.2	**83.0**	62.6	**92.8**	72.6	81.5	71.5	77.8	74.5
	YNU_wb	-	-	-	-	-	-	-	-	**78.8**

Table 1: Quantitative performance of models on 5 datasets. DE-All denotes the German BERT model trained on all German datasets, EN is the English BERT model trained on all English datasets, DE-All-multi is multilingual BERT trained on all German data, EN-multi is multilingual BERT trained on all English data, and Rules is the rule-based system trained on the train set corresponding to the test. The union of two or more models means classifying a text as offensive iff at least one of the models classifies it as offensive. Previously published top systems included for comparison are FHAC (Bornheim et al., 2021), ComMA (Kumar et al., 2020a), HASOCOne (Dowlagar and Mamidi, 2021), IIIT_DWD (Mishra et al., 2020), IIITK (Ghanghor et al., 2021), and YNU_wb (Wang et al., 2019).

One might attribute these annotations to the lack of hostile intent in these posts, but this would be in sharp contrast with FNen22† and FNen23†, which contain the same words, also lack any offensive content, but are nevertheless annotated as offensive (and profane in particular). The German sample, taken from the GermEval dataset containing longer Facebook comments, also contained several instances of sarcasm, which typically resulted in false negative predictions such as FNde4*†‡ and FNde5*†‡. Finally, the English sample contained several examples of data error, such as the inclusion of non-English text (FNen3†‡) or encoding issues (FNen13†‡).

8 Conclusion

We presented a set of experiments illustrating the potential of simple rule-based systems for high-precision detection of offensive text as an alternative or enhancement of deep learning methods, especially in low-resource scenarios where training data is only available from another language or another task/genre. We also presented manual error analysis on samples from two languages to indicate some challenging aspects that are characteristic either of the task, such as sarcasm, or the datasets, such as inconsistent annotations. We believe we have shown the potential of rule-based systems as simple, transparent, flexible solutions to text classification tasks in general, while at the same time we highlighted some issues of a popular task for which existing solutions are error-prone and sensitive to data bias.

ID	Text
FNen14†‡	How many people you planning to shag in September? — one person. the rest are a bonus https://t.co/FcS1FpxSvE
FNde1*†‡	@USER solch sinnfreie Beiträge... 🤦‍♀️🤦‍♀️🤦‍♀️
FPde2*	Schauspielen kann er nicht. Und inzwischen meint er, Ahnung von Allem zu haben. Schlimm dieser Typ
FPde4*	@USER...äh, Verzeihung! Fangen Sie doch einfach mal bei sich selbst, mit Ihren unnützen Motorrädern, an!
FNen1*†‡	@timesofindia How dare they call it Indian variant when they dint call it a #wuhanvirus or #chinesevirus?? India should file a legal case against WHO and China in international court.
FNen2*†‡	Sad reality of Indian news channels. A minute by minute coverage of elections while a common man struggles to find #covid treatment essentials. Useless News channels. #COVIDSecondWaveInIndia #CoronaPandemic #IndiaCovidCrisis #COVID19India #IndiaChoked #aajtak #zeenews #ABPnews
FPen1*‡	miya four creeps into every thought i have what the fuck
FPen2*†	@imtillyherron Happy MF birthday to my fave bitch out there!! thank you for always being YOU and for showing me that I shouldn't have to worry about what others might say thank you for being my motivation, my idol who radiates nothing but positive energ
FNen22†	Bitch I done did so much today I'm tired
FNen23†	would you fuck me? - ash — Idk who ash is? So you gotta tell me lol https://t.co/I0Jj7LNEho
FNde4*†‡	@USER Sie sind Hellseher?
FNde5*†‡	Oh....die Frau hat eine Glaskugel ? Ist ja interessant.
FNen3†‡	@ANI Naa desh ko corona se bachaya Naa WB elections jeeta itna campaigning ke baad Seriously Modi is big failure for India than what I thought. #ResignPMmodi
FNen13†‡	Windy says oh ya hoor sir . . . No long in. Shattered. Got myself a wee part time job. 3 days a month. First day. 12 hour shift. Bollocks ðŸ˜®ðŸ¤¦ðŸ »â€ â™‚ï¸ ðŸ¤£ Think Iâ€™ll give ma sel a 9/10 the day though. What an absolute fuking stonker eh ðŸ˜ŽðŸ”¥ðŸ™Œ

Table 2: Sample texts misclassified by any of our systems, grouped by error type. Text IDs indicate false positive (FP) or false negative (FN) and the models that made the false prediction. * denotes the language specific BERT model, † refers to the multilingual BERT model, ‡ marks the rule-based system.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2021. FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 105–111, Heinrich Heine University Düsseldorf.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for abusive language detection in english. *ArXiv*.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. 2018. Boolean decision rules via column generation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucia Donatelli, Meaghan Fowlie, Jonas Groschwitz, Alexander Koller, Matthias Lindemann, Mario Mina, and Pia Weißenhorn. 2019. Saarland at MRP 2019: Compositional parsing across all graphbanks. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 66–75, Hong Kong. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2021. Hasocone@fire-hasoc2020: Using bert and multilingual bert models for hate speech detection.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Kinga Gémes and Gábor Recski. 2021. TUW-Inf at GermEval2021: Rule-based and hybrid methods for detecting toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 69–75, Heinrich Heine University Düsseldorf, Germany.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Ritesh Kumar, Bornini Lahiri, Atul Kr. Ojha, and Akanksha Bansal. 2020a. ComMA@FIRE 2020: Exploring multilingual joint training across different classification tasks. In *FIRE*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020b. 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).

Kirti Kumari and Jyoti Singh. 2020. AI_ML_NIT_Patna @HASOC 2020: BERT models for hate speech identification in Indo-European languages. In *FIRE*, pages 319–324.

Piyawat Lertvittayakumjorn, Leshem Choshen, Eyal Shnarch, and Francesca Toni. 2021. GrASP: A library for extracting and exploring human-interpretable textual patterns.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Mandlia Chintak, 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.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR.

A. Mishra, Sunil Saumya, and Abhinav Kumar. 2020. IIIT_DWD@HASOC 2020: Identifying offensive content in Indo-European languages. In *FIRE*, pages 139–144.

Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3Idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in Indo-European languages. In *FIRE (Working Notes)*, pages 208–213.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive Tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

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.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. 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.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Amir Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

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, Düsseldorf, Germany.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Prithviraj Sen, Marina Danilevsky, Yunyao Li, Siddhartha Brahma, Matthias Boehm, Laura Chiticariu, and Rajasekar Krishnamurthy. 2020. Learning explainable linguistic expressions with neural inductive logic programming for sentence classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4211–4221, Online. Association for Computational Linguistics.

Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. 2019. HEIDL: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, Florence, Italy. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Serim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502. Association for Computational Linguistics.

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.

Bin Wang, Yunxia Ding, Shengyan Liu, , and Xiaobing Zhou. 2019. YNU_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. In *FIRE (Working Notes)*, pages 191–198.

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.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on

obstacles and solutions. *PeerJ Computer Science*, 7.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. 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). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Rule Systems

English HASOC 2021:

(fuck | asshole | whore | fucking | motherfucker | dick | bitch | useless | fuck-off | dick | shit | wank | bullshit | penis | bastard | shameless | fucker | piss-off | piss | clown)

$act \xrightarrow{prepagainst} country$

$shame \xrightarrow{ARG1} (media \mid person \mid publication \mid they \mid you \mid party \mid have \mid government)$

$shame \xrightarrow{ARG0} (vulture \mid elect \mid I \mid media \mid it \mid expose \mid you \mid have \mid obligate \mid support \mid nation \mid result \mid tell \mid person \mid get \mid vote \mid possible \mid religious \mid bastard \mid this \mid know \mid democracy \mid let \mid we \mid pull \mid and)$

$wanker \xrightarrow{mod} (.*)$

$embarrass \xrightarrow{ARG1} you$

$person \xrightarrow{mod} horrible$

$kill \xrightarrow{ARG1} person$

GermEval 2021:

(ARD | außer | ??? | silly | motherfucker | apparent | !!!. | foolishness | framing | arrogant | lach | asozial | council | arsch | bullshit)*

$propaganda \ \& \ NOT \ (.*) \xrightarrow{mod} propaganda$

$fact \ \& \ NOT \ (.*) \xleftarrow{mod} fact \xrightarrow{mod} (.*)$

$genericconcept \xrightarrow{op2} and$

$flüchtlinge \ \& \ NOT \ (.*) \xrightarrow{ARG1} flüchtlinge$

$(.*) \xrightarrow{.*} hashtag.* \ \& \ NOT \ (.*) \xrightarrow{ARG.*} hashtag.*$

$contrast \xleftarrow{op1} and \xrightarrow{mod} this$

$interrogative \xleftarrow{mode} and \xrightarrow{mod} only$

$(komm \mid willst \mid sell \mid heule \mid beschützt \mid schämen) \xrightarrow{ARG1} devour \xrightarrow{ARG1} you$

$(normal \mid level \mid complacency \mid issue \mid läuft \mid brechen \mid money \mid live \mid mention \mid leave \mid wahlzettel \mid thank \mid read \mid eingerechnet \mid car \mid prefer \mid außer \mid müssen) \xrightarrow{polarity} -$

English HASOC 2020:

(stupid | bitch | moron | hoe | damn | fuck | shit | ass | fucking | animal)

$rape \xrightarrow{manner} (.*)$

$rape \xrightarrow{ARG0} (.*)$

$rape \xrightarrow{ARG1} (.*)$

German HASOC 2020:

(hurensohn | maul | fresse | porno | abschaum | anal | ficken | fick)

$(müsste \mid neger \mid labern \mid dass \mid unerhört \mid scheiße \mid kelbt \mid performt \mid silly) \xrightarrow{mod} scheiße$

$(cause \mid wählen \mid richt \mid erzaehlt \mid say \mid muss) \xrightarrow{ARG1} scheiße$

$and \xrightarrow{op1} scheiße$

$(geblockt \mid ficken \mid write \mid tätscheln \mid strecken \mid können \mid spritzt \mid hope \mid need \mid \#ff \mid use \mid gefickt) \xrightarrow{ARG1} arsch$

$(kanacke \mid bekomm \mid star \mid lecken \mid thinspo \mid obacht \mid schickten \mid liebst \mid roxyred19589189 \mid pornos \mid cause \mid bluejeans \mid vergas \mid würde \mid geiler \mid fürn \mid full \mid hört \mid gefickt \mid fishnets \mid fisting \mid lutschen \mid geile \mid präsentiert \mid herangehensweise \mid hauen \mid peach \mid hift) \xrightarrow{mod} arsch$

$(grundeis \mid top \mid tail \mid genericconcept \mid verwöhnen \mid für) \xrightarrow{location} arsch$

$arsch \xrightarrow{mod} (euch \mid patriots \mid community \mid sammer \mid shoutoutxobella \mid \#rt \mid certain \mid dein \mid bineuerboss \mid plague \mid sexy \mid evt \mid fick \mid nackten \mid total \mid geil \mid nen \mid molligekleinesie83 \mid milf \mid stromkasteriks \mid otherwise \mid person \mid ariiiish \mid holgerewald1 \mid pornos \mid süsser \mid den \mid verdammten \mid cam \mid blöder \mid nix \mid genericconcept \mid privilegierten \mid meilenweit \mid würde \mid geiler \mid ersma \mid dorne \mid verfickten \mid dir \mid \#kostenlos \mid andy93893217 \mid erzähl \mid elektrogeräte \mid GT \mid veganhure \mid prallen \mid real \mid herrnewstime \mid mizunowaverider \mid kings \mid few \mid geilen \mid in \mid einzelfallinfos \mid voll \mid schön \mid pictimundi \mid gratis)$

English HASOC 2019:

(fucking | ass | bastard | rape | FuckTrump | fuck | vagina | dickhead | shithibbon | FatOrangeFuck | disgrace | shit)

$traitor \xrightarrow{ARG1} person$

$lie \xrightarrow{ARG0} you$

B Additional results

Test set	Train set	Offensive			Other			Average		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
GermEval2021	GermEval2021	67.3	19.4	30.2	66.5	94.4	78.1	66.9	56.9	61.5
GermEval2019	GermEval2019	71.3	53.6	61.2	80.5	89.9	84.9	75.9	71.7	73.8
	GermEval	63.0	70.6	66.6	85.3	80.4	82.8	74.1	75.5	74.8
	German HASOC	71.7	45.9	55.9	78.2	91.5	84.3	74.9	68.7	71.7
	German	73.0	63.9	68.2	84.0	88.9	86.4	78.5	76.4	77.4
	English multilingual	48.3	35.8	41.1	73.1	82.0	77.2	60.7	58.9	59.7
GermEval2018	GermEval2018	72.0	60.1	65.5	81.0	87.9	84.3	76.5	74.0	75.2
	GermEval	67.7	67.8	67.7	83.4	83.3	83.3	75.5	75.6	75.5
	German HASOC	77.2	39.4	52.1	75.0	94.0	83.4	76.1	66.7	71.1
	German	74.9	54.5	63.1	79.4	90.6	84.6	77.2	72.5	74.8
	English multilingual	53.8	22.7	31.9	69.3	90.0	78.3	61.6	56.3	58.8
German HASOC2020	German HASOC2020	69.6	74.7	72.0	89.2	86.5	87.8	79.4	80.6	80.0
German HASOC2019	German HASOC2019	33.2	77.9	46.6	94.4	70.2	80.5	63.8	74.1	68.5
	GermEval	32.0	83.1	46.2	95.4	66.4	78.3	63.7	74.7	68.8
	German HASOC	46.1	51.5	48.6	90.5	88.5	89.5	68.3	70.0	69.1
	German	35.5	73.5	47.8	93.7	74.5	83.0	64.6	74.0	69.0
	English multilingual	35.2	41.2	38.0	88.4	85.6	87.0	61.8	63.4	62.6
English HASOC2021	English HASOC2021	84.8	83.3	84.1	73.2	75.4	74.3	79.0	79.3	79.2
	GermanEval multilingual	77.8	18.9	30.4	40.5	91.1	56.1	59.2	55.0	57.0
	German HASOC multilingual	70.6	22.6	34.2	39.8	84.5	54.1	55.2	53.5	54.3
English HASOC2020	English HASOC2020	91.5	91.6	91.5	91.3	91.2	91.3	91.4	91.4	91.4
	GermanEval multilingual	66.9	12.3	20.7	51.0	93.8	66.0	58.9	53.0	55.8
	German HASOC multilingual	75.5	19.5	30.9	53.0	93.5	67.7	64.3	56.5	60.1
English HASOC2019	English HASOC2019	59.0	75.3	66.2	91.0	82.5	86.5	75.0	78.9	76.9
	GermanEval multilingual	51.0	34.4	41.1	80.3	89.0	84.4	65.7	61.7	63.6
	German HASOC multilingual	43.0	33.3	37.6	79.4	85.3	82.2	61.2	59.3	60.2

Table 3: All BERT models have been trained to optimize macro F1 on the validation set.

C Dataset sizes

Data	Offensive	Other
GermEval21 train	1122	2122
GermEval21 test	350	594
GermEval19 train	1287	2708
GermEval19 test	970	2061
GermEval18 train	1688	3321
GermEval18 test	1202	2330
German HASOC20 train	673	1700
German HASOC20 test	300	727
German HASOC19 train	407	3412
German HASOC19 test	136	714
English HASOC21 train	2501	1341
English HASOC21 test	798	483
English HASOC20 train	1856	1852
English HASOC20 test	807	785
English HASOC19 train	2261	3591
English HASOC19 test	288	865

D Models' performance on analyzed samples

Test	System	TP	TN	FP	FN	Prec	Rec	F1
	EN	64	23	11	2	85.3	97.0	90.8
	DE-multi	12	32	2	54	85.7	18.2	30.0
	Rules	32	32	2	34	94.1	48.5	64.0
EN	EN ∪ Rules	64	22	12	2	84.2	97.0	90.1
	DE-multi ∪ Rules	35	30	4	31	89.7	53.0	66.7
	EN ∪ DE-multi	64	22	12	2	84.2	97.0	90.1
	EN ∪ DE-multi ∪ Rules	64	21	13	2	83.1	97.0	89.5
	DE	12	62	5	21	70.6	36.4	48.0
	EN-multi	10	63	4	23	71.4	30.3	42.6
	Rules	4	66	1	29	80.0	12.1	21.1
DE	DE ∪ Rules	13	61	6	20	68.4	39.4	50.0
	EN-multi ∪ Rules	12	62	5	21	70.6	36.4	48.0
	DE ∪ EN-multi	15	58	9	18	62.5	45.5	52.6
	DE ∪ EN-multi ∪ Rules	16	57	10	17	61.5	48.5	54.2

Table 4: Quantiative performance of models on the samples used for error analysis

E Classifier errors on the analyzed samples

We list all false predictions made by any of the classifiers on either of the two samples used for manual analysis in Section 7, grouped by language, error type, and model. Table 5, 6, and 7 contain all false positive predictions on the English sample, Tables 8, 9, 10, and 11 group false negative predictions on the English sample by model. Table 12 contains false positive predictions on the German sample, Tables 13, 14, 15, and 16 contain false negative predictions on the German sample.

ID	Text
FP1*	miya four creeps into every thought i have what the fuck
FP2†	@imtillyherron Happy MF birthday to my fave bitch out there!! thank you for always being YOU and for showing me that I shouldn't have to worry about what others might say thank you for being my motivation, my idol who radiates nothing but positive energ
FP3	@mybmc Total mismanagement for 18-45 years. Fastest finger first. Slots are gone like hot cakes. BMC expects people to be on Twitter whole day. OTP system is pathetic. what is the need to log out if job is not done.Puts load on system unnecessarily. #VaccineRegistration #CovidVaccine
FP4	We should not unnecessarily curse PM. What are our people doing. Why they are not taking care of themselves. Even if you see conditions before 8 months in US and Italy. It was same. No conuntry has massive health infra to tackle so many cases. #ResignPMmodi
FP5	If I go 30 more minutes without some Samsung Virtual assistant pussy Iâ m gone fold ð ¯ https://t.co/t5lOT22eFz
FP6	Whatever is happening in Bengal is so disheartening to see! Being a Bong,it hurts more badly. Elections are conducted peacefully in every state except Bengal. Still can't figure out why there's always violence during polling. For God's sake,stop it! #BengalViolence #BengalBurning
FP7	Modi's super-spreader events went in vain. Mamata Banerjee's Trinamool Congress is on its way to form a government in West Bengal for the third time in a row. #IndiaCovidCrisis https://t.co/guehujzoD2
FP8	Indian journalists are being arrested, imprisoned, and tortured for speaking on the #COVID19 crisis in their country.Â India's use of security, defamation, and hate speech laws to detain critical voices is a grave violation of human rights. #ResignModi https://t.co/HzTGSOEtIN
FP9	I would have shot myself if I was the reason of so many deaths. #Resign_PM_Modi #ResignModi #andhbhakts https://t.co/hxkdyAQU3d
FP10	So we already know that WlKlPEDlA, funded by George S0r0s ONLY has misinformation about INDIA aNd HINDUISM but openly showing their intent of merging #Bengal with Bangladesh? People who opposed BJP rallies in #BengalElection2021 must know how dumb they were after #BengalBurning https://t.co/u9h18M36bs
FP11	@Dipankar_cpiml I, also lost my father. #Resign_PM_Modi

Table 5: False positive predictions of the English BERT system in the English HASOC 2021 dataset. Asterisks (*) denote examples that were also missed by the rule-based approach and the dagger (†) denotes the one also missed by the German multilingual BERT system.

ID	Text
FP12	During the times when the whole nation is facing and fighting the pandemic. Bengal is burning. The political parties sow the seeds (of hatred) and the innocents get to reap. #stopviolence #BengalBurning

Table 6: False positive prediction of the German multilingual BERT system on the English HASOC 2021 dataset.

ID	Text
FP13	#CovidVaccine #COVID19 if you have been injured or family member killed by vaccine ... you must report to fda and cdc. These vaccines must never get full approval. @cnn

Table 7: False positive prediction of the rule-based system on the English HASOC 2021 dataset because of the kill $\xrightarrow{ARG1}$ person relation.

ID	Text
FN1	@timesofindia How dare they call it Indian variant when they dint call it a #wuhanvirus or #chinesevirus?? India should file a legal case against WHO and China in international court.
FN2	Sad reality of Indian news channels. A minute by minute coverage of elections while a common man struggles to find #covid treatment essentials. Useless News channels. #COVIDSecondWaveInIndia #CoronaPandemic #IndiaCovidCrisis #COVID19India #IndiaChoked #aajtak #zeenews #ABPnews

Table 8: Offensive tweets in the English HASOC 2021 missed by all classifiers.

ID	Text
FN16	@ErLashe You have a fat pussy.
FN17	Drop it , shameless bully wood creatures .. run some awareness drive about #wuhanvirus /COVID 19 .. if u hv so much fan following (disgusting) than hopefully they wld listen to you .. #RadheTitleTrack not interested ..
FN18	@wahlstedt007 @pamela53blue Damn dude is being very nice! At a minimum he is at least a lying orange shitbag.....

Table 9: Offensive tweets in the English HASOC 2021 missed by the rule-based systems.

ID	Text
FN19	@smitabarooah Many shameless idiots in media want to project China positively as if nothing has happened since past 1year... #ChineseVirus destroyed many countries economy and killed millions... China should be punished
FN20	You think im tripping cause I need some dick. I'm tripping cause I want more than that. But everything goes over ya head!
FN21	i have shaved my dick, balls, armpits, ass, legs, face, head over the sink and in the shower and it sitll works absolutely fine
FN22	Bitch I done did so much today I'm tired
FN23	would you fuck me? - ash — Idk who ash is? So you gotta tell me lol https://t.co/I0Jj7LNEho
FN40	@narendramodi Have you no shame ? You count COVID meetings. You count election rallies. You count votes. While the country counts deaths. Ashamed to have voted for you. Ashamed to have been brainwashed like the whole country. Shame. #ResignModi #Resi
FN41	@PrathamShashank @LibtardsKaDoc @MamataOfficial @RahulGandhi BJP is not allowed in West Bengal so tell your modi to fuck off and never come back here #ResignModi
FN42	@Reuters The CCP Virus came from a Lab in Wuhan, China. China is ResponsibleÂ for the Spread ofÂ Corona VirusÂ Across theÂ World.ðŸŒ ðŸ¦ It's a Chinese VirusðŸ¤¬ðŸ¦ Fuck You ChinaðŸ–•ðŸ‡¨ðŸ‡³ðŸ–•ðŸ¦ #CKMKB #ChuttadChinesePartyÂ CCP #ChineseVirus
FN43	@mePankajBhaskar @derekobrienmp Hey #AndhBhakt your supremo #ResignModi preferred dancing in elections to visiting hospitals. No oxygen, no beds, no ventilators just elections... that turned the nation into graveyard. Shame #ResignPMmodi , Didi o didi #
FN44	Blatant example of many among one double speak of Modi government. #ModiKaVaccineJumla is real Jumla for Bankers. They have been deliberately left to die. When it is work everyone becomes master of Bank but in case of welfare measures all gets mouth paralysis. Shame @DFS_India https://t.co/6VrCI5dlkA
FN45	@BSYBJP @mla_sudhakar @nimmasuresh @RAshokaBJP @narendramodi @AmitShah @ShobhaBJP @DVSadanandGowda What is your take on this. There won't be any fume without fire. Still you people want to play with the people's life? Shame on you people #Resig
FN46	@WrittenByHanna To hell with being the bigger person. Ima be an asshole instead. https://t.co/7f4q2J3syW
FN47	@Rute_Cauze @farsazadi Lmao mate could you be any more of a wanker ðŸ˜
FN48	Bitch better be thanking God Almighty because he couldve had his ass seriously BEAT TO A BLOODY pulp https://t.co/XrmEiQBIxq
FN49	How dare you motherfucker trying to tell everyone to be cool ? R u fucking serious ? #StopAsianHate https://t.co/lRNzlYOtcD
FN50	@MattHancock This may all be true. But... What did you do to piss off big Dom?
FN51	@OkamotoSan @XIA0SVIA @kobrastarships @IzzieNai @klancezstarz "gatekeeper" say that to my face, fucker. SAY THAT TO MY FACE, INSTEAD OF BLOCKING ME, COME ON.
FN52	@PatMcAfeeShow @TWooodley @jakepaul Double leg that fucker and drop hellbows
FN53	'Trace of blood' motherfucker if that's a trace of hate to see what this guy makes of periods 💀
FN54	@LAVLUXX Nawww.... let me tell y'all what kinda dick this really is. This dat " Ima Barely wet the head dick and then shove it in to wake her up dick."
FN55	@VFLBrentford_ @JxshCFCx @TheVFL_ The next day the fucker was dead https://t.co/3K6uAKjrjr
FN56	I love people that believe they 'deserve' to be addressed a certain way and that you should walk on egg shells even when they're being a jackass because I'm gonna burst the fuck outta that self-entitled bubble.
FN57	WTF is happening in our country #BengalBurning #SpinlessBJP https://t.co/SU3CAT87ri

Table 10: Offensive tweets in the English HASOC 2021 missed by the German multilingual BERT-based classifier.

ID	Text
FN3	@ANI Naa desh ko corona se bachaya Naa WB elections jeeta itna campaigning ke baad Seriously Modi is big failure for India than what I thought. #ResignPMmodi
FN4	@narendramodi @drharshvardhan It was already knew that second wave would come. You should have kept necessary stocks. You people are so dumb, unaccountable and ignorant. Feels like curse to have leaders like you.U Resign! #ModiMustResign #ModiKaVaccineJu
FN5	@@thecoastguy â€¦..But is a real Wanker, you know.
FN6	glad I prefer Dutch beer to Italian beer anyway. Cunt.
FN7	Day after day, whatever little this regime does gets progressively more frustrating and infuriating. This sheer level of incompetence when people literally die on the streets. You lot can't really shake off what you truly are. Murderous genocidal bigots. #ResignPMmodi https://t.co/Mcwi1jIklI
FN8	I as a citizen feel ashamed of you being the PM of this country. PLEASE REALISE YOUR POWER AND SAVE PPL, YOU ARE NO MORE A CHAIWALA BUT A BLOODY PM of this country. Or just leave the status and go run a tea stall ð ¡ #ResignPMmodi #ModiResign #ShameOnModi #ShameOnYou 3/3
FN9	So how has @narendramodi thought of vaccination to rural areas ? Online registration ? Really ? Then he will have to give them network first ,then cell phones And then after few More years these people will get vaccine for sure . What a fool India has chosen to lead #ResignModi
FN10	@mollierose0827 Sorry but ur not worth anymore than a hooker on the side of the street
FN11	Geezus you got your old age #CovidVaccine already, first I think! Now you berate renters, a gigantic swathe of younger people who canâ€™t do what you say? Ignorant & boring. https://t.co/rtB-nqYWRtM
FN12	No strategy for vaccination. Worried how this halfwit govt shall do it. #ModiKaVaccineJumla
FN13	Windy says oh ya hoor sir . . . No long in. Shattered. Got myself a wee part time job. 3 days a month. First day. 12 hour shift. Bollocks ðŸ˜®ðŸ¦ðŸ »â€ â™‚ï¸ ðŸ¤£ Think Iâ€™ll give ma sel a 9/10 the day though. What an absolute fuking stonker eh ðŸ˜ŽðŸ”¥ðŸ™Œ
FN14	How many people you planning to shag in September? — one person. the rest are a bonus https://t.co/FcS1FpxSvE
FN15	not exclusive..... lol Ange just given him permission to shag her daughter unwittingly? & Louie LoveDay from #Hollyoaks just NOOOO lol lol #HolbyCity
FN24	@spectatorindex Because of Worst PM & Planning #ResignModi
FN25	@BJP4India @sambitswaraj You and your party should be ashamed...Thoda Sharm Kar lo Bhai #Resign_PM_Modi
FN26	@BJP4India @sambitswaraj You and your party should be ashamed...Thoda Sharm Kar lo Bhai #Resign_PM_Modi
FN27	@rocktredblood @abbyeddymom @mikecanjerk @MorpheusDeZion @wash77033 @IamWendums @Sammy5586 @briansjelks @cdhthegreat @RocketSilasEra @jtkilla_13 Iâ€™m not a fan of Capela. He talked all that smack and whimped out when the chips were down. Harden shouldâ€™
FN28	@HootHoottooh Your timeline is filled with hatred & bigotry towards Republicans. Stay Woke HootyHoo!
FN29	People in India are still dying like files!! Our govt is not helping at all. The world is seeing the failure and this will go down in history bold and bloody! Stop #CentralVistaProject! #ModiKaVaccineJumla #modioxygendo #ModiDisasterForIndia #ModiResign
FN30	Scientists are begging for data but #Modi is going against science and taking no advice. Global community can't help us if we hide data and deny science. #ResignPMmodi please, you're incompetent. #AndhBhakt #ModiKaVaccineJumla #ModiResign https://t.co/tilRANVPdg
FN31	Reality of Bengal, Comparing the TMC with Nazis, the BJP dubbed the govt in Bengal as fascist. BJP-IT Cell,trying it's best like always, spreading hate & misleading, manipulating everyone. Their manipulator in chief @amitmalviya is not stopping. #CruelMamata #BengalBurning https://t.co/HJMp24Qev4
FN32	@airnewsalerts . @TOIIndiaNews @ANI @BBCHindi @nytimes Please Stop fooling people, have been checking your pathetic portal every two hours and no slot available.. for proof view the screen shot attached.!! #ModiResign #ModiKaVaccineJumla #ModiResignOrRepe
FN33	#IndiaCovidCrisis #IndiaNeedsOxygen Shameful of BJP for not being able to manage health system even after so many weeks; they slept since the 1st pandemic despite knowing the second wave would be dangerous and oxygen is key to live/ Delhi is a death bed all over !!
FN34	@AfreenFatima136 @HayaatK58831077 To cover up your #Hinduphobia and vulturism, donâ€™t take help of #islamophobia. We have #Criminalsphobia only. Those who insult dead are with criminal behaviour. And no one cares for you fellows. Get Out
FN35	Total failure of health infrastructure„, , low economy...death of democracy #ExitModi #ResignPMmodi https://t.co/Wt8nLu9kJz
FN36	If you can't handle me at my "show me on the doll where the feminist hurt you", then you probably deserve me at my "get fucked you incel shitbag".
FN37	@vampirebritcom wanker omg yes
FN38	Bloody hell that got closer than I would have liked!
FN39	@Baddiel That's is ' I'm going to Shag you' look.

Table 11: Offensive tweets in the English HASOC 2021 sample missed by the German multilingual BERT and rule-based systems.

ID	Text
FP1	Alle die hier so böswillig kommentieren, denkt doch einfach mal an eure eigene Jugend. Was hat sich da besonders eingeprägt? Die Klassenfahrt an die Ostsee mit den Klassenkameraden, die Ferienreise wo ihr eine neue Freundschaften geschlossen habt, das Auslandssemester wo ihr die erste Liebe getroffen habt, der Abend wo ihr beim Tanz eure/n Frau / Mann kennengelernt habt. Erinnerungen die mit Lachen und einem Gefühl der Unbeschwertheit gefüllt sind. Das lässt sich durch nichts ersetzen! Und man sollte der Jugend zumindest zugestehen, das es schwer fällt darauf zu verzichten. Und natürlich ich das nicht so schlimm wie einen Menschen zu verlieren (ich hab selbst meinen Vater dieses Jahr verloren), aber es ist trotzdem unglaublich frustrierend.
FP2	Schauspielen kann er nicht. Und inzwischen meint er, Ahnung von Allem zu haben. Schlimm dieser Typ
FP3	Nach Ihrer gestrigen Sendung steigt bei mir die Angst, dass es zu diesem ”Unfall” kommt, vor allem dieser Mister Weinberg macht mir riesig Angst, er könnte für eine Mehrheit gesprochen haben/sprechen 🤔😔 ! Wir werden sehen, wie das mit der Gerechtigkeit/ Demokratie in den USA bestellt ist. Gute Gesundheit.
FP4	@USER...äh, Verzeihung! Fangen Sie doch einfach mal bei sich selbst, mit Ihren unnützen Motorrädern, an!
FP5	Für Europa wird sich nichts ändern, die alten Männer und die Sümpfe hinter ihnen wird die ”America First” Politik fortführen. Europa sollte endlich zusammenrücken und mit einer demokratisch gewählten Stimme auftreten!
FP6*	@USER Dass es schon immer so war, ist erwiesenermaßen falsch. Als die Demokraten noch für Sklaverei eintraten, waren die Republikaner für die Abschaffung dieser. Erst mit Harding als Präsident wandelten sich die Republikaner zu einer konservativen Partei.
FP7*	@USER Die Entwicklungen in den USA sind für uns alle evident, nicht nur für uns deutsche Bürger. Trump’s Verhalten stärkt nicht nur in den USA die rechten Ränder, sondern auch bei uns. Deutschland/Europa muss aufpassen, nicht zu sehr in diesen Sog gezogen zu werden. Unsere Regierungen müssen sich klar gegen die Politik Trumps wenden, wenn es um die Zukunft der NATO und der UN geht. Trump steht für Rassismus, Klassenkampf, Frauenfeindlichkeit, Homophobie und Umweltverschmutzung. All das geht letztendlich auch mich an.
FP8*	Sie durfte / konnte leider nicht so viel sagen.. Lag am Weinberg 😉 .. Hat auch einen super tollen Akzent sehr sympathisch ✌
FPde9*	@USER , es kommt sicher auch auf die Wahl in den USA an. Mit Trump als Präsident kann sich Putin mehr erlauben. Dann juckt es ihn natürlich nicht,was Europa sagt. Wenn der komplette Westen Sanktionen verhängt, sieht es wohl anders aus.
FP10†	Österreich macht es vor! Ich wünsche mir eine solche Talk Gruppe im ARD und ZDF: https://www.servustv.com/videos/aa-2549wnudn2112/?fbclid=IwAR1X69rdRKZ8epwJTVXNGM1inpdsSXgtYrxkbGqucid9XFSA4IQPEX1vINc Außerdem wünsche ich, dass auch unsere Gesundheitsbehörden mal endlich die Wahrheit sprechen: https://www.facebook.com/jorg.achatz.52/videos/3813957285300040/?extid=Q6XnDuTDc0XGrZ5r

Table 12: False positive predictions on the GermEval 2021 toxic dataset. Examples without notation are missed by the German BERT, asterisks (*) denote examples missed by the rule-based approach and the dagger (†) denotes the ones missed by the English multilingual BERT system.

ID	Text
FN1*†	@USER solch sinnfreie Beiträge... 🤦‍♀️🤦‍♀️🤦‍♀️
FN2*†	@USER, sie merken nicht mal wie sie manipuliert werden...
FN3*†	@USER welcher amerikanische Präsident war in KEINEN Krieg verwickelt?
FN4*†	@USER Sie sind Hellseher?
FN5*†	Oh....die Frau hat eine Glaskugel ? Ist ja interessant.
FN6*†	Niemanden in den USA interessiert es, was irgendwer in Deutschland über die Wahl redet. Worin liegt für mich der Mehrwert dieser Sendung ?
FN7*†	@USER es wird keine andere Meinung zugelassen
FN8*†	Die Frage ist ja nicht welche Meinung man teilt. Egal ob Trump oder Biden verlieren wird wieder die Unterschicht..
FN9*†	Trolle…
FN10†	@USER zum Glück werden Kommentare nicht vorgelesen sonst wäre ihrer auch ein Teil davon 🤔
FN12*†	@MEDIUM, „*innen“ ist links-ideologischer Gender-Sprech und hat im öffentlich-rechtlichen Rundfunk nichts verloren!
FN13*†	@USER Soviel ich weiß, hat die SED und ihre Nachfolgeorganisation PDS heute „die Linke“ Gerichtsfest festgestellt unter und mit Gysi als Vorsitzender,damals hunderte von Millionen DM „Parteivermögen“ verschoben. Ist mir jetzt zu langweilig und doof all die Berichte und Artikel heraus zu suchen für dich. Wer das nicht weiß, verweigert sich einfach der Realität.
FN14*†	Ach ja... alles doch so einfach... für 2 Cent mehr CO2 neutral tanken 😸
FN15*†	Das Format @MEDIUM gehört abgeschafft. - KOMPLETT.
FN16*†	@USER 1950er? eher 1650. Ein immer größerer Teil der US-Gesellschaft ist hinter die Aufklärung zurück gefallen und kommt mit ihrem Rassismus und ihrer Bigotterie daher, als sei sie gerade erst von Bord der Mayflower gestiegen.
FN17*†	@USER Wind, Wasser, Sonne, Kuhdung, Straßenbau etc Sie reden und schreiben ja auch nur und haben nichts zu sagen.
FN18*†	@USER Spekulativ, unfundiert, frustgeschwängert und auch noch ohne jedwede Erwähnung der Alternative, die Biden gegenüber steht...aah Facebook, right.
FN19*†	@USER Das zeigt mal wieder ein ganz intelligenten Beitrag
FN20*	Die US-Wahlen 2020 Das Horoskop des Präsidentschaftskandidaten And the winner is - Joe Biden Wahlforscher, lasst uns doch nicht länger leiden – ins Amt kommt bitteschön Joe Biden! Selten war das Ergebnis der Auseinandersetzung um die Präsidentschaft der Vereinigten Staaten so eindeutig vorhersehbar wie 2020 – jedenfalls aus kosmischer Sicht. Obgleich sich der demokratische Herausforderer während des gesamten Wahljahres nicht unbedingt gemocht fühlt, ja sogar einsam und melancholisch, so hat er dennoch einen Trumpf in der Hand, der in der Öffentlichkeit sticht: er hat die seriöseren, die glaubwürdigeren, vor allem die verlässlicheren Argumente! Und die werden zunehmend den Ausschlag geben. Joe Biden wird ab dem 10. Oktober bis zum Wahltag immer mehr argumentative Punkte sammeln. Zudem bietet sein Horoskop eine Projektionsfläche, die kollektive Sehnsüchte anzieht wie duftender Lavendel die Bienen. Selbst, wenn diese Wünsche anschließend nicht alle erfüllt werden. Diese Konstellation ähnelt der des Wahlkämpfers Helmut Kohl, der im Juli 1990 von den blühenden Landschaften im Osten schwärmte. Es ist, als sehnte sich die Mehrheit der Amerikaner nach der Ära eines missratenen Republikaners wieder nach einem Heilsbringer, nach einem Präsidenten, der für alle da ist. Die Bevölkerung spürt, das da wirklich einer ist, der den Vereinigten Staaten dienen will. Und das gibt definitiv am Dienstag, dem 03. November, den Ausschlag! © Jörg Vos, 28.09.2020
FN21*	Trump ist ein pathologischer Lügner. Diese Krankheit heißt: „Pseudologia phantastica“. Häufiger wird heute der Begriff pathologisches Lügen verwendet. Eine besondere Form der Pseudologia phantastica stellt das Münchhausen-Syndrom dar, bei dem der Patient körperliche Beschwerden erfindet und durch Lügen untermauert, um Aufmerksamkeit zu bekommen. In der modernen psychiatrischen Klassifikation wird diese Störung unter „andere Persönlichkeits- und Verhaltensstörungen/artifizielle Störung“ (ICD-10: F68.1) eingeordnet.
FN22*	Erneut haben Zehntausende Menschen in der belarussischen Hauptstadt Minsk gegen Präsident Alexander Lukaschenko demonstriert. Ohne Masken oder Sicherheitsabstand. Und das schon seit Wochen. Die Corona Fallzahlen steigen aber nicht. Wenn Demos in Deutschland gegen die Corona Maßnahmen stattfinden, wird Nazi oder Verschwörung gerufen und Abstand oder fehlende Masken angemahnt. Die selben Kritiker feiern die Demos ohne Sicherheitsmaßnahmen von Belarus als ein MUSS. Alles LUG und BETRUG.

Table 13: False negative predictions made by the German BERT system in the GermEval 2021 dataset. Asterisks (*) mark examples missed by the rule-based approach and the dagger (†) denotes the ones missed by the English multilingual BERT system.

ID	Text
FN11	@USER Sie haben also Beweise, dass die Russen so dumm sind, dass sie einen bekannten Oppositionellen mit einem russischen Gift ””anfixen”” (er überlebt es), ihn dann nach Russland einfliegen, um ihn zu ””behandeln”” (er überlebt es) und ihn dann in die Hände der deutschen Russlandfeinde geben, damit diese eindeutig das Gift feststellen können? Da bekommt man ja schon beim Hören und Lesen einen hochroten Kopf vom Fremdschämen!

Table 14: False negative prediction made by the English multilingual BERT system on the GermEval 2021 dataset.

ID	Text
FN23	Ich würde zu gerne wissen, womit Frau Baerbock im Winter heizt. Oder hängt sie sich stilecht im Tippi nur ein Bärenfell um (ein Co2 neutrales)
FN24	Es ist wie mit Corona. Argumente zählen nicht mehr und Wissenschaft ist böse. In den USA greift der Wahnsinn in Form von Evangelikalen um sich, bei uns nennen sich die Realitätsverweigerer nur anders.
FN25	Nee nee was für ein Gelaber! Sorry - das brauch ich nicht für meine Gebühren! Was wirklich wichtig ist weiß die junge Frau nicht! Stereotype einfach nachplappert! So schaut’s aus! Zefix!
FN26	Die finden immer wieder einige Verräter 🤣🤣🤣

Table 15: False negative predictions made by the rule-based and the English multilingual BERT systems.

ID	Text
FN27	Die Baerbock hat schon oft bewiesen, dass sie nicht gerade die hellste Kerze auf der Torte ist. Sie hat nur Kobolde im Kopf
FN28	Dummes Geschwätz. Er sitzt alle Dumnscheätzer aus. Zu Recht
FN29	@USER Auf was beziehen sie ihren Beitrag? Ich meine, ich habe Merkel nicht erwähnt, ich habe keine anderen Politiker oder deren Parteien in Deutschland erwähnt, ich Urteile hier nicht über Amerika... 🤦 Irgendwie ist ihr Beitrag komplett aus der Luft gerissen! Und ja, es gibt auch Leute, die radikal Merkels Meinung vertreten und eine andere nicht zulassen... was Mutti Merkel macht und sagt, ist richtig, obwohl es nicht richtig ist. Auch diese Leute gibt’s, wenn sie das hören wollten... 🤦‍♂️🤷‍♂️
FN30	@USER nur weil du zu faul zum selber suchen bist? 🙈 😂
FN31	@MEDIUM danke für die Info. Aber ist es nicht so , dass alles bisherige was die Herren Lauterbach u Drosten usw von sich gaben u geben spekulativ u geschätzt ist und Gottseidank nichts davon so eingetroffen ist . Trotzdem dürfen sie weiter ihre Spekulationen öffentlich von sich geben. Laden Sie doch mal zb einen Prof Bhakdi ein u.lassen Sie Leute wie ihn Rede u Antwort stehen. Da hätten sie unglaubliche Einschaltquoten!!! Und nicht nur einseitige spekulative Redner...

Table 16: False negative predictions made by the rule-based system.