Offensive text detection across languages and datasets using rule-based and hybrid methods

Kinga Gémes[1,2], Ádám Kovács[1,2] and Gábor Recski[1]

[1] TU Wien, Favoritenstraße 9-11., Vienna, 1040, Austria
[2] Budapest University of Technology and Economics, Műegyetem rkp. 3., Budapest, H-1111, Hungary

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 quantitative and qualitative comparison of their performance with deep learning models on 5 datasets across multiple languages and shared tasks.

Keywords

offensive text, rule-based methods, human in the loop learning

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 definitions 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 using five datasets of English and German social media text. 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 main contributions of the paper are the following:

- A rule-based method for offensive text detection using semantic parsing and graph patterns
- 5 high-precision rule systems for English and German offensive text detection based on datasets from two shared tasks
- Quantitative evaluation of our rule systems, deep learning baselines, and their ensembles across 5 datasets, demonstrating that rule based and hybrid systems can outperform deep learning models in cross-dataset and cross-language settings.
- Detailed error analysis of each system on samples of 100 posts each from one English and one German dataset.

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://github.com/GKingA/offensive_text.

CIKM'22: Advances in Interpretable Machine Learning and Artificial Intelligence Workshop, Oct 17–21, 2022, Atlanta, GA

kinga.gemes@tuwien.ac.at (K. Gémes); adam.kovacs@tuwien.ac.at (´. Kovács); gabor.recski@tuwien.ac.at (G. Recski)
0000-0003-0626-9644 (K. Gémes); 0000-0001-6132-7144 (´. Kovács); 0000-0001-5551-3100 (G. Recski)

2. Related Work

Datasets As pointed out already in a 2017 survey [1], 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 [2] 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 [3] and GermEval [4]. 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 [5] and 2020 [6] use task definitions similar to GermEval. Other widely used datasets with a narrower scope include the data provided by the TRAC [7, 8] and HatEval [9] 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 [10] and contextual embeddings [8, 11]. As in many popular text classification tasks, the most widely used neural language models are based on the Transformer architecture [12], and in particular BERT-based models [13] are the basis of the state of the art machine learning systems for most datasets, including the best-performing systems on GermEval2021 [14], GermEval2019 [15], HASOC 2020 English [16] and HASOC 2020 German [8, 17]. Top systems enhance quantitative performance by optimizing metaparameters such as maximum sentence length or number of training epochs [18, 19], by training on joint subtask labels [20] or utilizing multiple Transformer based models to counteract the small dataset sizes [14], by pretraining on additional hate speech corpora [21], training jointly on different corpora [8], or by using adversarial learning [22]. Further deep learning methods used in offensive text detection include LSTMs [23, 24], CNNs [25, 26], or both [27], sentence embeddings [28], and ensembles of multiple machine learning models [27, 29].

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. [30], [31], and [32] for the controversy about using attention weights as explanation). Yet even the more mature methods for interpreting neural networks (e.g. LIME [33]) 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 [34] and integer programming [35] to the training of probabilistic grammars over semantic graphs [36]. Human-in-the-loop (HITL) approaches involve generating rule candidates to be reviewed by experts, e.g. by extracting textual patterns [37] or semantic structures [38]. Rule-based approaches are also often combined with ML methods, e.g. by incorporating lexical features into DL architectures [39, 40] or voting between rule-based and ML systems [41, 42, 43].

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. We choose two recent tasks that use identical labeling schemes and also have one language in common (German) allowing us to perform various cross-dataset experiments. Our experiments involve datasets in German and English only, these are the two languages for which we are able to build rule systems and also perform qualitative analysis (see Section 7) in addition to quantitative results, allowing us to investigate the ability of both ML and rule-based models to transfer between tasks as well as languages.

The GermEval shared task was organized in 2018 [44], 2019 [45], and 2021 [4]. German Twitter posts were annotated for the 2018 and 2019 challenges, the 2021 task used comments from a news-related Facebook group. 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*.

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 [46], 2020 [47], and 2021 [3]. 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.

Our rule-based solutions are built using POTATO[1] [48], a framework that enables the rapid construction of graph-based rule systems and has recently been used for text classification in multiple domains and languages. Input text is parsed into Abstract Meaning Representations (AMR, [49]), directed graphs of concepts representing the semantics of each sentence. For English texts we use a pretrained Transformer-based AMR parser [50] and the `amrlib`[2] library, for German we construct AMRs from text using a multilingual, transition-based system [51] via the `amr-eager-multilingual`[3] library. A rule system for a task consists of lists of patterns over graph representations of text for each possible class, and a text is predicted to belong to a given class iff at least one pattern in the class's list of patterns matches the corresponding graph. Graphs must be directed and can be edge- and/or node-labeled. Individual patterns are directed 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 as a single boolean statement in disjunctive normal form (DNF) of boolean predicates corresponding to graph patterns, in this regard the method is similar to the approach of [35] and [34] (see Section 2).

To construct rule systems efficiently, POTATO implements 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 systems manually. We shall describe the individual rule systems built for our experiments in Section 5.

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 between English and German data, by training models using multilingual BERT on datasets

[1] https://github.com/adaamko/POTATO

[2] https://amrlib.readthedocs.io/en/latest/

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

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* 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` for German and `bert-base-uncased` for English) as well as the multilingual model (`bert-base-multilingual-cased`). 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. Hyper-parameters 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*,[5] to compensate for unbalanced labels, as suggested by [52]. 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.

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.

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*. 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 is available from our repository.

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.

HASOC organizers argue that using macro-average F1-score counteracts class imbalance [46]. We follow this practice in our evaluation, especially since many of the top participating systems do not publish scores for individual classes. Our main results on the test portions of each of the 5 corpora is presented in Table 1. On each dataset we 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. We also present the scores of the top-performing system for each dataset.

The DL models trained on data from the same task achieve the best results. 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. Rule systems achieve the highest precision values

[4] The dependencies and BERT models are noted in our repository
[5] https://scikit-learn.org/

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
	DE-GermEval2021	67.3	19.4	30.2	66.5	94.4	78.1	66.9	56.9	61.5
	EN-all-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-all-multi ∪ Rules	54.9	27.4	36.6	67.0	86.7	75.6	60.9	57.1	58.9
	DE-all ∪ EN-all-multi	62.3	44.9	52.2	72.1	84.0	77.6	67.2	64.4	65.8
	DE-all ∪ EN-all-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**
	DE-HASOC2020	69.6	74.7	72.0	89.2	86.5	87.8	79.4	80.6	80.0
	EN-all-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-all-multi ∪ Rules	62.1	61.7	61.9	84.2	84.5	84.3	73.2	73.1	73.1
	DE-all ∪ EN-all-multi	51.1	94.7	66.4	96.6	62.6	76.0	73.8	78.6	76.2
	DE-all ∪ EN-all-multi ∪ Rules	51.2	**95.0**	66.5	**96.8**	62.6	76.0	74.0	78.8	76.3
	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-all	80.3	95.2	**87.2**	88.7	61.5	**72.6**	**84.5**	**78.4**	81.3
	EN-HASOC2021	84.8	83.3	84.1	73.2	75.4	74.3	79.0	79.3	79.2
	DE-all-multi	82.7	23.9	37.1	42.2	**91.7**	57.8	62.4	57.8	60.0
	DE-GermEval-multi	77.8	18.9	30.4	40.5	91.1	56.1	59.2	55.0	57.0
	DE-HASOC-multi	70.6	22.6	34.2	39.8	84.5	54.1	55.2	53.5	54.3
	EN-all ∪ 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-all ∪ DE-all-multi	79.3	95.5	86.6	88.8	58.8	70.7	84.0	77.1	80.4
	EN-all ∪ DE-all-multi ∪ Rules	78.8	**95.7**	86.4	89.1	57.3	69.8	83.9	76.5	80.1
	NLP-CIC	-	-	-	-	-	-	-	-	**83.1**
EN HASOC2020	Rules	**95.3**	74.6	83.7	78.6	**96.2**	86.5	86.9	85.4	86.2
	EN-all	90.2	90.5	**90.3**	90.2	89.9	**90.1**	**90.2**	**90.2**	90.2
	EN-HASOC2020	91.5	91.6	91.5	91.3	91.2	91.3	91.4	91.4	91.4
	DE-all-multi	79.3	20.9	33.1	53.7	94.4	68.5	66.5	57.7	61.8
	DE-GermEval-multi	66.9	12.3	20.7	51.0	93.8	66.0	58.9	53.0	55.8
	DE-HASOC-multi	75.5	19.5	30.9	53.0	93.5	67.7	64.3	56.5	60.1
	EN-all ∪ 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-all ∪ DE-all-multi	86.6	91.9	89.2	91.2	85.4	88.2	88.9	88.6	88.8
	EN-all ∪ DE-all-multi ∪ Rules	86.0	**92.3**	89.1	**91.5**	84.6	87.9	88.7	88.5	88.6
	IIITK	-	-	-	-	-	-	-	-	**93**
EN HASOC2019	Rules	**73.2**	35.1	47.4	81.6	**95.7**	**88.1**	**77.4**	65.4	70.9
	EN-all	59.6	76.7	**67.1**	91.4	82.7	86.8	75.5	**79.7**	77.5
	EN-HASOC2019	59.0	75.3	66.2	91.0	82.5	86.5	75.0	78.9	76.9
	DE-all-multi	53.1	47.9	50.4	83.2	85.9	84.5	68.1	66.9	67.5
	DE-GermEval-multi	51.0	34.4	41.1	80.3	89.0	84.4	65.7	61.7	63.6
	DE-HASOC-multi	43.0	33.3	37.6	79.4	85.3	82.2	61.2	59.3	60.2
	EN-all ∪ 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-all ∪ DE-all-multi	51.5	82.6	63.5	**92.8**	74.1	82.4	72.1	78.4	75.1
	EN-all ∪ 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. The language codes are DE for German and EN for English. The 'all' denotes the language specific BERT model trained on all datasets of that language, 'all-multi' is multilingual BERT trained on all language specific 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 [14], ComMA [8], HASOCOne [17], IIIT_DWD [24], IIITK [16], and YNU_wb [23]. The NLP-CIC team, whose system was reported by shared task organizers to have achieved the highest F1 score on the shared task [3], did not publish a description of their methods, and is only included for the sake of completeness.

on each dataset, which 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 tasks and 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 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%.

7. Error Analysis

In this 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. The quantitative results on this sample are noted in the README of our repository. 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 our repository.

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. 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†‡).

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

[1] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in:

Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[2] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021). URL: https://peerj.com/articles/cs-598/. doi:https://doi.org/10.7717/peerj-cs.598.

[3] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, 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, FIRE 2021, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–3. URL: https://doi.org/10.1145/3503162.3503176.

[4] J. Risch, A. Stoll, L. Wilms, M. Wiegand, 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, Association for Computational Linguistics, Düsseldorf, Germany, 2021, pp. 1–12. URL: https://aclanthology.org/2021.germeval-1.1. doi:10.48415/2021/fhw5-x128.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. URL: https://aclanthology.org/S19-2010. doi:10.18653/v1/S19-2010.

[6] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Çağrı Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188.

[7] ws-2018-trolling, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018. URL: https://aclanthology.org/W18-4400.

[8] R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, ComMA@FIRE 2020: Exploring multilingual joint training across different classification tasks, in: FIRE, CEUR, Hyderabad, India, 2020, pp. 823–828. URL: http://ceur-ws.org/Vol-2826/T10-3.pdf.

[9] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[10] P. Chiril, F. Benamara Zitoune, V. Moriceau, M. Coulomb-Gully, A. Kumar, 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, ATALA, Toulouse, France, 2019, pp. 351–360. URL: https://aclanthology.org/2019.jeptalnrecital-court.21.

[11] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5838–5844. URL: https://aclanthology.org/2020.emnlp-main.470. doi:10.18653/v1/2020.emnlp-main.470.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., Long Beach, CA, USA, 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. arXiv:1706.03762.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. of NAACL, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] T. Bornheim, N. Grieger, S. Bialonski, 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, Association for Computational Linguistics, Heinrich Heine University Düsseldorf, Germany, 2021, pp. 105–111. URL: https://aclanthology.org/2021.germeval-1.16.

[15] A. Paraschiv, D.-C. Cercel, 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), German Society for Computational Linguistics & Language Technology, Erlangen, Germany, 2019, pp. 398–404.

[16] N. Ghanghor, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, 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, Association for Computational Linguistics, Kyiv, 2021, pp. 197–203. URL: https://aclanthology.org/2021.ltedi-1.30.

[17] S. Dowlagar, R. Mamidi, Hasocone@fire-hasoc2020: Using bert and multilingual bert models for hate speech detection, 2021. URL: https://arxiv.org/pdf/2101.09007.pdf. arXiv:2101.09007.

[18] K. Kumari, J. Singh, AI_ML_NIT_Patna @HASOC 2020: BERT models for hate speech identification in Indo-European languages, in: FIRE, CEUR, Hyderabad, India, 2020, pp. 319–324. URL: http://ceur-ws.org/Vol-2826/T2-29.pdf.

[19] P. Liu, W. Li, L. Zou, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 87–91. URL: https://aclanthology.org/S19-2011. doi:10.18653/v1/S19-2011.

[20] S. Mishra, S. Mishra, 3Idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in Indo-European languages, in: FIRE (Working Notes), CEUR, Kolkata, India, 2019, pp. 208–213. URL: http://ceur-ws.org/Vol-2517/T3-4.pdf.

[21] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in english, in: Proceedings of the 5th Workshop on Online Abuse and Harms, volume Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Bangkok, Thailand, 2021, pp. 17–25. URL: https://aclanthology.org/2021.woah-1.3. doi:10.18653/v1/2021.woah-1.3.

[22] T. Tran, Y. Hu, C. Hu, K. Yen, F. Tan, K. Lee, S. Park, HABERTOR: An efficient and effective deep hatespeech detector, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7486–7502. URL: https://aclanthology.org/2020.emnlp-main.606. doi:10.18653/v1/2020.emnlp-main.606.

[23] B. Wang, Y. Ding, S. Liu, , X. Zhou, YNU_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language, in: FIRE (Working Notes), CEUR, Kolkata, India, 2019, pp. 191–198. URL: http://ceur-ws.org/Vol-2517/T3-2.pdf.

[24] A. Mishra, S. Saumya, A. Kumar, IIIT_DWD@HASOC 2020: Identifying offensive content in Indo-European languages, in: FIRE, CEUR, Hyderabad, India, 2020, pp. 139–144. URL: http://ceur-ws.org/Vol-2826/T2-5.pdf.

[25] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 85–90. URL: https://aclanthology.org/W17-3013. doi:10.18653/v1/W17-3013.

[26] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on Twitter, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 41–45. URL: https://aclanthology.org/W17-3006. doi:10.18653/v1/W17-3006.

[27] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 759–760. URL: https://doi.org/10.1145/3041021.3054223. doi:10.1145/3041021.3054223.

[28] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, V. Varma, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 70–74. URL: https://aclanthology.org/S19-2009. doi:10.18653/v1/S19-2009.

[29] A. Nikolov, V. Radivchev, Nikolov-radivchev at SemEval-2019 task 6: Offensive Tweet classification with BERT and ensembles, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 691–695. URL: https://aclanthology.org/S19-2123. doi:10.18653/v1/S19-2123.

[30] S. Serrano, N. A. Smith, Is Attention Interpretable?, in: Proceedings of the 57th Annual Meeting of the

Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951. URL: https://aclanthology.org/P19-1282. doi:10.18653/v1/P19-1282.

[31] S. Wiegreffe, Y. Pinter, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20. URL: https://aclanthology.org/D19-1002. doi:10.18653/v1/D19-1002.

[32] S. Jain, B. C. Wallace, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. URL: https://aclanthology.org/N19-1357. doi:10.18653/v1/N19-1357.

[33] M. T. Ribeiro, S. Singh, C. Guestrin, "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, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[34] P. Sen, M. Danilevsky, Y. Li, S. Brahma, M. Boehm, L. Chiticariu, R. Krishnamurthy, 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), Association for Computational Linguistics, Online, 2020, pp. 4211–4221. URL: https://www.aclweb.org/anthology/2020.emnlp-main.345. doi:10.18653/v1/2020.emnlp-main.345.

[35] S. Dash, O. Gunluk, D. Wei, Boolean decision rules via column generation, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., Montréal, Canada, 2018. URL: https://proceedings.neurips.cc/paper/2018/file/743394beff4b1282ba735e5e3723ed74-Paper.pdf.

[36] L. Donatelli, M. Fowlie, J. Groschwitz, A. Koller, M. Lindemann, M. Mina, P. Weißenhorn, 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, Association for Computational Linguistics, Hong Kong, 2019, pp. 66–75. URL: https://aclanthology.org/K19-2006. doi:10.18653/v1/K19-2006.

[37] P. Lertvittayakumjorn, L. Choshen, E. Shnarch, F. Toni, GrASP: A library for extracting and exploring human-interpretable textual patterns, https://arxiv.org/abs/2104.03958, 2021. arXiv:2104.03958.

[38] P. Sen, Y. Li, E. Kandogan, Y. Yang, W. Lasecki, 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, Association for Computational Linguistics, Florence, Italy, 2019, pp. 135–140. URL: https://www.aclweb.org/anthology/P19-3023. doi:10.18653/v1/P19-3023.

[39] A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti, HurtBERT: Incorporating lexical features with BERT for the detection of abusive language, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 34–43. URL: https://aclanthology.org/2020.alw-1.5. doi:10.18653/v1/2020.alw-1.5.

[40] E. W. Pamungkas, V. Patti, 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, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://aclanthology.org/P19-2051. doi:10.18653/v1/P19-2051.

[41] A. Razavi, D. Inkpen, S. Uritsky, S. Matwin, Offensive language detection using multi-level classification, in: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 16–27. URL: http://www.eiti.uottawa.ca/~diana/publications/Flame_Final.pdf. doi:10.1007/978-3-642-13059-5_5.

[42] K. Gémes, G. Recski, 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, Association for Computational Linguistics, Heinrich Heine University Düsseldorf, Germany, 2021, pp. 69–75. URL: https://aclanthology.org/2021.germeval-1.10.

[43] K. Gémes, A. Kovács, M. Reichel, G. Recski, Offensive text detection on English Twitter with deep learning models and rule-based systems, in: FIRE 2021 Working Notes, CEUR, Gandhinagar, India, 2021, pp. 283–296. URL: http://ceur-ws.org/Vol-3159/T1-29.pdf.

[44] M. Wiegand, M. Siegel, J. Ruppenhofer, 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, Austrian Academy of Sciences, Vienna, Austria, 2018, pp. 1–10. URL: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-84935.

[45] J. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, 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, German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, München, Germany, 2019, pp. 352–363. URL: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/GermEvalSharedTask2019Iggsa.pdf.

[46] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, M. Chintak, A. Patel, 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, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[47] T. Mandl, S. Modha, A. K. M, B. R. Chakravarthi, 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, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[48] Á. Kovács, K. Gémes, E. Iklódi, G. Recski, Potato: explainable information extraction framework, in: Proceedings of the 31st ACM International Conference on Information and Knowledge Management, CIKM '22, Association for Computing Machinery, 2022, p. 4897–4901. URL: https://doi.org/10.1145/3511808.3557196. doi:10.1145/3511808.3557196.

[49] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract Meaning Representation for sembanking, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 178–186. URL: https://www.aclweb.org/anthology/W13-2322.

[50] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[51] M. Damonte, S. B. Cohen, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1146–1155. URL: https://aclanthology.org/N18-1104. doi:10.18653/v1/N18-1104.

[52] G. King, L. Zeng, Logistic regression in rare events data, Political Analysis 9 (2001) 137–163. URL: https://www.cambridge.org/core/journals/political-analysis/article/logistic-regression-in-rare-events-data/1E09F0F36F89DF12A823130FDF0DA462. doi:10.1093/oxfordjournals.pan.a004868.