Offensive text detection on English Twitter with deep learning models and rule-based systems

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

[1]*TU Wien*
[2]*Budapest University of Technology and Economics, Department of Automation and Applied Informatics*

Abstract
This paper describes the systems the TUW-Inf team submitted for the HASOC 2021 shared task on identifying offensive comments in social media. Besides a simple BERT-based classifier that achieved one of the highest F-scores on the binary classification task, we also build a high-precision rule-based classifier using a custom framework for human-in-the-loop learning. Both of our approaches are also evaluated qualitatively by manual analysis of 150 tweets, which also highlights possible controversies in the ground truth labels of the HASOC dataset.

Keywords
social media data, hate speech detection, rule-based methods, deep learning, text classification

1. Introduction

This paper describes the systems submitted to the HASOC 2021 shared task on identifying offensive comments in social media. We experimented with standard, well performing deep learning (DL) architectures and an explainable, rule-based system built semi-automatically using a custom framework. We developed our models for the English dataset and submitted runs for both the binary and fine-grained classification tasks. Our best system is a simple combination of these solutions. We also perform manual, qualitative analysis of the labels predicted by our two independent systems on a sample of 150 tweets to illustrate the strengths and weaknesses of both approaches. The paper is structured as follows. An overview of recent related work is provided in Section 2 and our methods are described in Section 3. Quantitative evaluation is in Section 4, Section 5 presents our qualitative analysis, and Section 6 concludes the paper. All software described in the paper is publicly available under an MIT license at https://github.com/GKingA/tuw-inf-hasoc2021.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

kinga.gemes@tuwien.ac.at (K. Gémes); adam.kovacs@tuwien.ac.at (Á. Kovács); mx.markus.rei@gmx.net (M. Reichel); gabor.recski@tuwien.ac.at (G. Recski)

2. Related work

2.1. Tasks and datasets

The growing interest in automatically detecting offensive text, especially in social media, has led to the creation of multiple tasks and datasets in recent years. Overlapping task definitions have used terms including *hate speech detection*, *toxicity detetction*, and *offensive text detection*, and recent shared tasks on these topics include Semeval [1, 2], GermEval [3, 4, 5], and HASOC [6, 7]. The HASOC and GermEval'18 and '19 subtasks define the challenges of detecting offensive language as well as classifying them into fine-grained categories, the labels used include *profane*, *abusive* or *hateful*, and *insulting* or *offensive*. The Semeval tasks, both dubbed *OffensEval*, contain three subtasks, a binary classification between offensive and non-offensive text, the distinction between targeted and non-targeted offensive texts, and the identification of the target (individual/group/other). Based on the definitions of each challenge, this latter task of identifying whether the target of offensive speech is an individual or a specific group roughly corresponds to the GermEval/HASOC distinctions between *insulting/offensive* and *abusive/hateful*, while non-targeted offensive texts are similar to the *profane* category. Tasks and datasets are available for a growing number of languages. HASOC has covered English, German, Hindi, and Marathi languages, while OffensEval provides datasets for Arabic, Danish, English, Greek, and Turkish.

2.2. Approaches

Regarding the approaches to this group of tasks, leaderboards ranking systems by quantitative performance are dominated by models based on the Transformer architecture [8], most prominently BERT [9] and similar pretrained models. In addition to fine-tuning these standard models on the training data available for a given task, top systems improve quantitative performance by optimizing metaparameters such as maximum sentence length or number of training epochs [10, 11], by training on joint subtask labels [12], by pre-training on additional hate speech corpora [13], or by using adversarial learning [14]. Yet another competitive approach is an ensemble model [15] combining the outputs of a wide range of supervised learning architectures including BERT, recurrent neural networks (RNN), multi-layer perceptrons (MLP), support vector machines (SVM), etc. The best performing system on the HASOC 2020 English task used GloVe embeddings and a long short-term memory (LSTM) model to achieve 51.52% F1-score on the binary classification subtask [16]. In our work we also explore rule-based methods that offer high explainability and configurability of classifier systems. Approaches that involve explainable representations include a DL-based extractor of surface features such as ngrams of words and parts-of-speech (POS) [17] and an architecture for detecting online harassment by the pattern-based identification of offensive text linked to references of a person [18, 19].

2.3. Explainable AI

State-of-the art neural models with millions of parameters, such as the BERT-based classifier described in Section 3.1, are difficult to interpret and explain by nature. Various techniques exist for explaining black-box models [20, 21], one of the most popular is to explain a prediction by visualizing attention weights over the words of the input and treat it as a possible interpretation of the entire model [22, 23]. Several recent studies have examined the validity of attention as a source of explanation/interpretation of deep learning models [24, 25, 26]. *Is Attention Interpretable?* [24] concludes that the highest attention weights often fail to have an impact and when used as a ranking of importance it fails to explain model decisions. *Attention is not not Explanation* [25] argues that attention weights should be interpreted as one of several possible explanations of a model.

Because of the black-box nature of neural models and increasing concerns about their interpretability and trustworthiness, explainable AI (XAI) [27] methods have been the subject of growing interest. Rule-based models are interpretable and explainable, but they tend to be fragile, and constructing them manually requires considerable effort and domain expertise. Recent work on (semi-)automatic rule learning includes RuleNN [28], a neural network architecture for sentence classification that learns first-order logic rules over shallow semantic representations, and human-in-the-loop (HITL) machine learning systems [29, 30] that use labeled data to extract rule candidates that can be manually updated by domain experts. The framework we use for building a rule-based solution to the HASOC shared task is most similar to these HITL approaches and will be described in Section 3.2.

3. Method

This section presents our DL-based and rule-based systems developed for the HASOC 2021 task. Our main focus was the binary classification problem, but we also briefly describe the systems used for the fine-grained task. Both approaches will be evaluated in Section 4, while the qualitative analysis in Section 5 will explore their advantages and disadvantages.

3.1. Supervised learning

For training we use the datasets provided by HASOC 2019 [6], 2020 [7] and 2021 [31]. Text preprocessing involved removing hashtag symbols, replacing emoticons with their textual representation using the *emoji*[1] Python library, and substituting currencies and urls with special tags using the regex based library *clean-text*[2]. Finally, we use our own regular expressions for masking usernames with the *[USER]* tag.

We train the model *bert-base-uncased* with a single linear classification layer. We used Adam optimizer with a weight decay value of 10^{-5} and initial learning rate of 10^{-5}. Batch size was set to 8 and each model was trained for 10 epochs to determine the optimal number of iterations (2) based on the macro F-score measured on the validation set. We shuffle the training data after every iteration. Because of the distribution of the training data we used the balanced weighted loss function of *sklearn*[3], inspired by [32]. Metaparameters were optimized using a portion of the training set held out for validation but included in training the models for our final submission. We did not use the test portions of the 2019 and 2020 HASOC datasets for either training or development, these are used for the qualitative analysis in Section 5. This architecture was used

[1] https://pypi.org/project/emoji/

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

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

for both subtasks, for the multi-class dataset we trained a binary classifier for each of the three fine-grained categories and at prediction time choosing the label with the highest likelihood according to the corresponding classifier.

For the fine-grained classification task (1b) we also experimented with a Random Forest classifier using interpretable features such as word ngrams and edges of AMR graphs as described in Section 3.2. Ngram features were extracted after lowercasing and stopword removal, for $n \leq 3$, the most frequent 2500 features were used for training the classifier, with default metaparameters of the *sklearn* implementation[4] and class weights set to correspond to the ratio of labels in the dataset.

3.2. Rule-based classification

We use a custom framework for human-in-the-loop (HITL) learning of features over semantic representations of tweets. We construct Abstract Meaning Representation (AMR, [33]) graphs from text using Transformers' T5 encoder [34] and the amrlib[5] Python library. Sentences are encoded as sets of binary features corresponding to AMR subgraphs of at most 2 edges, keeping only the 2500 most frequent features. Our framework then trains a standard decision tree and ranks features based on their gini importance. Top features are presented to the features as candidates for patterns that trigger the *HOF* label and the user can build a set of rules by accepting, rejecting, or modifying these, while continuously having access to the predictions of each rule and the compiled rule set on the training data. Rules can be specified as node- and edge-labeled subgraphs, and node labels can be replaced with regular expressions (regexes). For example, the rule $\textit{wanker} \xrightarrow{\textit{mod}} \textit{(.*)}$ matches subgraphs where the word *wanker* is attached to any other word with the *mod* edge. If such a rule causes many false positives, the system provides the option of automatically refining it by ranking the subgraphs defined by the regex based on their performance on the training data. When thus refining the rule $\textit{shame} \xrightarrow{ARG1} \textit{(.*)}$, which in itself would yield 43 false positives for the 202 true positives, and including only subgraphs with a precision of at least 90% and a recall of at least 1%, we constructed the rule $\textit{shame} \xrightarrow{ARG1} \textit{(media|person|publication|they|you|party|have|government)}$, which yields 8 false positives for 103 true positives. We used this framework to build a simple rule set for the binary subtask. Besides AMR subgraphs with a single node corresponding to single words such as *fuck, whore, dick*, etc., some larger AMR subgraphs were also extracted. The complete rule set used for our submission is the following:

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

$\textit{act} \xrightarrow{\textit{prepagainst}} \textit{country}$

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

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

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

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

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

$\textit{embarrass} \xrightarrow{\textit{ARG1}} \textit{you}$

$\textit{person} \xrightarrow{\textit{mod}} \textit{horrible}$

$\textit{kill} \xrightarrow{\textit{ARG1}} \textit{person}$

The framework can be used for any classification task and with any graph representation of text and will be described in more detail in a forthcoming paper.

3.3. Ensemble

Finally, we implement two simple ensemble methods for combining the outputs of our two independent approaches. On the binary task, the *union* method predicts a text to be hateful or offensive (*HOF*) if at least one of our two systems predicts it, while the *logreg voting* trains a logistic regression model with two features, the probability of the *HOF* label output by the BERT model and the binary output of the rule-based system.

4. Quantitative results

We participated in the first subtask of the 2021 HASOC challenge [35], and provided solutions only for the English language dataset. Subtask 1A involves binary classification of Tweets, where the labels *HOF* and *NOT* correspond to *Hate and Offensive* and *Non Hate-Offensive*. Subtask 1b is the fine-grained classification of tweets into *Hate speech* (*HATE*), *Offensive* (*OFFN*), and *Profane* (*PRFN*) tweets. We train our DL system using the training data provided as well as the training portions of the datasets provided for for the 2019 and 2020 shared tasks [6, 7]. The rule-based system was developed using the 2021 dataset only.

Official test results on the binary classification task is presented in Table 2 (results on the fine-grained task are omitted for lack of space but are available in [31]). Our BERT-based method achieves a competitive F1-score, placing third among 56 systems on the *HOF* class and ninth when measured by the average F1-score across the two classes. On the *HOF* class our rule-based system has the highest precision among all systems and the union of the two systems achieves the highest recall. This is expected, since the rules are designed for high precision and are independent of the BERT system. Results of a simple ablation study measuring the effect of each improvement over the standard BERT-based classifier is presented in Table 1. For the binary task we also evaluated our systems on the 2019 and 2020 test sets, since these were used for the qualitative analysis in Section 5, these figures are presented in Table 3.

5. Qualitative analysis

We examined the output of both types of systems on a randomly selected sample of 150 tweets from the 2019 and 2020 test sets (the gold labels for the 2021 test were only released shortly before the submission deadline). For both approaches we used the models with the highest

Table 1
Effect of individual improvements over the standard BERT-based classifier.

	Offensive			Not offensive			Average		
	P	R	F	P	R	F	P	R	F
BERT + both	80.34	95.24	87.16	88.66	61.49	72.62	84.50	78.36	79.89
BERT + weighted loss	85.41	84.34	84.87	74.65	76.19	75.41	80.03	80.26	80.14
BERT + preprocessing	80.15	95.61	87.20	89.36	60.87	72.41	84.75	78.24	79.81
BERT	82.46	88.35	85.30	78.17	68.94	73.27	80.31	78.64	79.28

Table 2
Official test results on Task 1a

	Offensive			Not offensive			Average		
	P	R	F	P	R	F	P	R	F
NLP-CIC (top)	85.11	90.98	87.95	83.17	73.00	78.15	84.14	81.99	83.05
TUW Logreg voting	81.31	93.23	86.87	85.25	64.60	73.50	83.28	78.91	80.18
TUW BERT-based	80.34	95.24	87.16	88.66	61.49	72.62	84.50	78.36	79.89
TUW Union voting	79.81	95.61	87.00	89.23	60.04	71.78	84.52	77.83	79.39
TUW Rule-based	87.17	45.11	59.45	49.54	89.03	63.66	68.35	67.07	61.56

Table 3
Results on the 2019 and 2020 binary task

	Offensive			Not offensive			Average		
	P	R	F	P	R	F	P	R	F
BERT-based	80.5	86.8	83.6	90.8	86.1	88.4	85.7	86.5	86.0
Rules 2019, 2020	91.3	64.2	75.4	80.2	95.9	87.3	85.7	80.1	81.4
Rules 2021	91.9	53.2	67.4	75.7	96.9	85.0	83.8	75.0	76.2
Union (BERT + Rules 19-20)	79.2	87.4	83.1	91.0	84.7	87.8	85.1	86.1	85.4
Union (BERT + Rules 21)	79.9	87.6	83.4	91.2	85.2	88.1	85.4	86.4	85.7

performance on the 2019 and 2020 validation sets. For the BERT-based approach we used the model trained on all training data between 2019 and 2021, while for the rule-based approach we used two separate rule sets for the 2019 and 2020 samples, each built based only on the training data from the respective year. The 2019 rule set contains 12 keywords and 2 AMR edges:

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

$\textit{traitor} \xrightarrow{\textit{ARG1}} \textit{person}$

$\textit{lie} \xrightarrow{\textit{ARG0}} \textit{you}$

The 2020 rule set contains 10 keywords and 3 (underspecified) AMR edges (see Section 3.2 for details):

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

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

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

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

The goal of our analysis was not only to observe the nature of errors made by each system, but also to better understand the task as it is implicitly defined by the gold labels in the dataset. In a similar analysis that we performed in earlier work [36] we examined equal-size samples of each of the four types of data points based on their predicted and gold labels (true positive, false positive, true negative, and false negative). Since here we also wish to compare the performance of two systems, we ran both systems on a sample of 150 sentences and manually examined all predictions.

The sample contains 85 tweets with the gold label *NONE*, while the number of tweets annotated as profane (*PRFN*), offensive (*OFFN*), and hateful *HATE* is 37, 15, and 13, respectively, and in this analysis we focus on the coarse-grained, binary classification task of deciding whether a tweet is labeled as one of these three classes or *NONE*. All tweets that were misclassified by at least one of the two systems are listed below. The rule-based system made only one false positive prediction (FP1), while the BERT system had 13 false positive errors (FP1-13), these are listed in Table 4. 10 tweets were falsely labeled as non-offensive by both systems (FN1-10, Table 5), and 14 additional tweets were missed only by the rule-based system (FN11-24, Table 6). The remaining 113 tweets in the sample were correctly classified by both systems (72 as non-offensive and 41 as offensive), these were also examined as part of our analysis and the full sample is available from our repository[6]. We do not consider ourselves more qualified to judge whether tweets are offensive than the annotators participating in the creation of the dataset, our goal is to further our understanding of whether a rule-based systems' decisions may be less controversial due to efforts to avoid obvious bias such as learning data artefacts, and whether this may suggest new methods of quantitative and/or qualitative evaluation for the task of detecting offensive text.

We believe the most typical case among the 13 false positive predictions in Table 4 are those which have likely been classified as non-offensive because of their sensitive topic and/or the presence of words or phrases typical of offensive tweets (FP3, FP4, FP5, FP6) (although it is worth noting that the word *damn*, which may make the status of FP3 controversial, warranted the *PRFN* label in the case of FN10). There are also several tweets whose status as non-offensive we believe to be more controversial, in particular FP2 and FP8, which explicitly call one or more persons *stupid* and *dumb*, respectively. The 10 tweets that were falsely classified as non-offensive by both systems (Table 5) include some examples of offensive text without using offensive words (FN1, FN2, FN3), but also some where the offensive status is only apparent when the URL is followed to access the video attached to the tweet (FN4, FN8), and a couple tweets where the source of the label is entirely unclear to the authors (FN6, FN9). As for the 14 offensive tweets that were only missed by the rule system (Table 6), these once again contain many clear examples of offensive text (FN12, FN16, FN19, etc.) but also some tweets whose status may be questioned despite their labels (e.g. FN13, FN15, FN17). Furthermore, 4 of these 14 tweets may be considered errors in the data, since two are at least partially non-English (FN14, FN18) and two are clipped so that the offensive content is missing from the dataset (FN15, FN20).

[6] https://github.com/GKingA/tuw-inf-hasoc2021/blob/main/error_analysis/sample.tsv

Table 4
False positive predictions of the BERT-based systems. The line marked with an asterisk (*) indicates the single tweet that was also falsely classified as offensive by the rule-based system

ID	Text
FP1*	Yeah, so, Islam is an idea, not a race. ...a terrible, hateful, idea and you disgrace yourself in its defense. Everybody likes the legal immigrants. Most of them are voting Trump. :) Sincerely, some guy, native of some place.
FP2	Definition of FOOL as per http://dictionary.com , http://en.oxforddictionaries.com “a silly or stupid person; a person who lacks judgment or sense / a clown” Why @nipfp_org_in @bsindia etc are tolerating him?
FP3	@ConnorJaundrell So damn gooood
FP4	If a Hindu mob were to lynch Muslim doctors, every political leader, Modi to Mamata to Rahul to Nitish to Kejri would announce a slew of compensation, visit the victims, and do elaborate Man ki Baats. Mujhe Maro, Mera Doctors ko maat maro, maat maro mere bhaiyon! Hindus disowned
FP5	Mamata Banerjee totally lost Plot in dealing with the Doctors. Immediate action could have prevented crisis. But BJP and Bhakts should refrain from commenting on Mamata’s Statements. In similar situation, BJP MOS wanted protesting doctors to become Naxals so that they can be shot
FP6	An image of arrested sex racketeers is going viral with different false claims. One of the false claim says insinuates that ’Hindu terrorists’ were caught in Allahabad. Another false claim says that Rohingya Muslims were kidnapping boys in MP... https://timesofindia.indiatimes.com/news/fake-alert-photo-of-sex-racket-gang-shared-as-hindu-terrorists-caught-in-allahabad/articleshow/70469604.cms?utm_medium=referral&utm_campaign=iOS-app&utm_source=facebook.com …
FP7	Did you know that Syed Ali Shah Geelani had given a shutdown call yesterday (Friday) in support of so called ’political prisoners’? Not a single shop was shut in Srinagar. Life was normal in Kashmir. Infact traffic jam in many areas. Thugs of Hurriyat can’t fool Kashmiris now.
FP8	The Twitter troll army has gone so mad that they’re uninstalling Zomato to show their support knowing that they’ll install it again after few days when lured by offers and discounts. I want to ask dumb Twitterians isn’t this hypocrisy?
FP9	No Muslim MP from your party comes. Who is keeping them behind? You. There’s a difference b/wtheir words&ideology. Narasimha Rao was responsible for Babri Masjid demolition, despite being PM he couldn’t do anything. Now there’s Modi who wants to work on his ideology. @asadowaisi
FP10	RT @xctlot: ”yeah, I keep my AR by the nightstand” ”Omg you monster” ”In case I need to shoot a fetus” ” 😍😍😍 ” https://t.co/J3RNe3FSZF
FP11	Ye hain ’Bankey Bihari’ , ye 19 July ko hi bhartee huay hain TROLL ARMY mein .... Inki bhasha dekhiye , Maa baap se bhi ye shayad aise hi baat karta hoga .. Kaash ye shakha ki jaga school gaya hota ! May Maa Durga ease his pain of not achieving anything except TROLLSHIP in life !
FP12	Muslim 2 his wife Talaq-Talaq-Talaq,wife chats back Modi-Modi Modi.This is a Victory 4 Fair sex.India rejoices today as da Triple Talaq Bill Passed Jai Hind
FP13	whenever I said about modi ji bhakts be like #boycottzomato #Zomato #IStandWithAmit (iss pappu koi pata batayega jisne ye amit wala trend start kiya?) #CCDMemories

6. Conclusion

We have presented two approaches to the task of detecting and classifying offensive English tweets. On the binary classification task our standard BERT-based classifier achieved the third highest F-score on the *HOF* class and the ninth highest average F-score across the *HOF* and *NOT* classes, among 56 teams. We also used a custom framework for human-in-the-loop learning of graph features to construct rule systems over Abstract Meaning Representations of tweets,

Table 5
Offensive tweets missed by both systems.

ID	Cat.	Text
FN1	HATE	world Muslim population explosion is real & will permanently change course of human civilization. If BJP not take it biggest challenge n control the population, Dark days r Ahead. What has not happened in last 10K yrs will happen in next 50 years! @AmitShah @narendramodi
FN2	HATE	Yeah especially when there will be 2 Muslim political party in the state against one secular party BJP.
FN3	HATE	That's not white supremacy though is it. That means whites thinking AND ACTING as though they are superior to others. What these Jew/Muslim joint actions are concerned with is protecting stuff like Halal/Kosher, circumcision and the right to wear religious garb. 1/2
FN4	HATE	@BernieSanders Bernie will never be the President of the United States of America. https://t.co/O0FPo0bTc9
FN5	OFFN	Course you are mate! Just keep still while I buckle the cuffs of your jacket together behind your back #fucktrump Peace&MuchLve
FN6	OFFN	RT @JenniThorburn: It's pure privilege to watch #Eurovision and ignore Palestinian people. You're emboldening an apartheid regime. You're…
FN7	OFFN	RT @jen_jennnnnnnn: jake paul: you're very mean and full of hatred cody ko and noel miller: https://t.co/3cOau8o2E2
FN8	OFFN	RT @NiteLoops: Kim Kardashian and Kanye West naming their child Psalm West made me think of this video https://t.co/9ouWog1GEO
FN9	OFFN	Best Totichor in the worldproud you sir ❤️❤️❤️
FN10	PRFN	RT @animatrocities: My sister doesn't have a twitter, she's only 14 but dammmmn she's gonna go far!! https://t.co/N16K1VsT7i

which can be used either as standalone high-precision classifiers or as a supplement to increase the recall of our BERT-based method. A detailed qualitative analysis on a sample of 150 tweets was presented to investigate the nature of errors made by each of the two systems and also as an illustration of possible controversies in the dataset's gold labels.

Acknowledgments

Research conducted in collaboration with Botium GmbH.

Table 6
Offensive tweets missed by the rule-based system only

FN11	HATE	Maybe whilst there, you could have taken questions from journalists, rather than hiding & releasing propoganda bulletins. If you’re a serious PM, you would allow some scrutiny without No.10 spin jockeys pumping out silly soundbites. Until then the nickname stays: #BackDoorBoris
FN12	HATE	@SkyNews Wherever they go their moronic brain dead fans wreak havoc.
FN13	HATE	#boycottzomato #IStandWithAmit #Zomato you have two rules simultaneously for hindus and for muslims..... And that’s not fair.... We live in a democratic secular country where every 1 has the right of choice.... By hurting the majority sentiments, you have taken a risk...Not fair.
FN14	HATE	Madam ji pls also raise your loud voice against the exodus of Kashmiri pandits by your so called ’ ’ who raped,butchered,acid attacked hindu women just because of their religion..pls also raise your voice against the slogan’kashmir mein rehna hai toh Allah hu akbar
FN15	OFFN	U no what, i cant b mad at paul p... i mean the celtics retired this man’s number an he won wit them. But ryan smh…https://t.co/LYDCzzytt3
FN16	OFFN	He’s always been a fraud, cheat & racist. But this behavior is off the charts insane. #FuckTrump
FN17	OFFN	RT @staypufd: He’s also one who said they had proof of collusion!!! He’s a liar all around, just like Schiff! https://t.co/3znTbXt1cq
FN18	OFFN	Halala was never part of islam lekin halala ke naam par mulle khoob maje le rabe tumahri maa behan beti ke aur tum bhi maje se dekh rahe unki izzat lutate huye. Aise to baat baat par bomd fod dete ho lekin ek bhi muslim londa inn mullon ke against kuch nahi bolta. Dhikkar hai
FN19	PRFN	Wonder if its Mamata/Yesudas tweet or use of ’marketing fuck up’ that helpd me lose 15 folowrs yest. Wazzat abt separatng wheat from chaff?
FN20	PRFN	Ass usual seculars... thoughts? Big hearted opinions??? https://t.co/6ViYU01L6K
FN21	PRFN	Good economy my ass! One of our Lowes this morning eliminated 2 departments and fired 6 people without notice. One was there 11 years. #lowes #fucktrump
FN22	PRFN	@DrCoolRanch Damn.
FN23	PRFN	RT @kolokomiks: should i pursue a relationship with someone who doesn’t watch game of thrones? — why the fuck should that be a criteria htt…
FN24	PRFN	Basically Mamata just tags any protest against her as whatever her enemy of the day is. The BJP did try hijacking the protest but NRS doctors themselves told them to fuck off.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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, Düsseldorf, Germany, 2021, pp. 1–12. URL: https://netlibrary.aau.at/urn:nbn:at:at-ubk:3-798. doi:10.48415/2021/fhw5-x128.

[6] 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. 1417. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[7] 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. 2932. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[8] 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., 2017, pp. 5998–6008. URL: http://papers.nips.cc/

paper/7181-attention-is-all-you-need.pdf. arXiv:1706.03762.

[9] 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.

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

[11] 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.

[12] 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), 2019, pp. 208–213. URL: http://ceur-ws.org/Vol-2517/T3-4.pdf.

[13] T. Caselli, V. Basile, J. Mitrovi, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in english, ArXiv (2020). arXiv:2010.12472.

[14] 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, 2020, pp. 7486–7502. URL: https://aclanthology.org/2020.emnlp-main.606. doi:10.18653/v1/2020.emnlp-main.606.

[15] 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.

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

[17] Z. Zhang, L. Luo, Hate speech detection: A solved problem? The challenging case of long tail on Twitter, Semantic Web 10 (2019) 925–945. URL: https://doi.org/10.3233/SW-180338. doi:10.3233/SW-180338.

[18] U. Bretschneider, T. Wöhner, Detecting online harassment in social networks, in: Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems, ICIS 2014, Auckland, New Zealand, December 14-17, 2014, Association for Information Systems, Auckland, New Zealand, 2014. URL: https://aisel.aisnet.org/icis2014/proceedings/ConferenceTheme/2.

[19] U. Bretschneider, R. Peters, Detecting offensive statements towards foreigners in social media, in: Proceedings of the 50th Hawaii International Conference on System Sciences, 2017, pp. 2213–2222. URL: http://hdl.handle.net/10125/41423. doi:10.24251/HICSS.2017.268.

[20] R. Ghaeini, X. Fern, P. Tadepalli, Interpreting recurrent and attention-based neural models:

a case study on natural language inference, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4952–4957. URL: https://www.aclweb.org/anthology/D18-1537. doi:10.18653/v1/D18-1537.

[21] 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. 11351144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[22] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 606–615. URL: https://www.aclweb.org/anthology/D16-1058. doi:10.18653/v1/D16-1058.

[23] J. Lee, J.-H. Shin, J.-S. Kim, Interactive visualization and manipulation of attention-based neural machine translation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 121–126. URL: https://www.aclweb.org/anthology/D17-2021. doi:10.18653/v1/D17-2021.

[24] 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.

[25] 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.

[26] 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.

[27] D. Gunning, Explainable artificial intelligence (XAI), 2016. URL: https://www.darpa.mil/program/explainable-artificial-intelligence.

[28] 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.

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

[30] 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.

[31] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech, in: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[32] 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.

[33] 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.

[34] 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.

[35] 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, CEUR, 2021. URL: http://ceur-ws.org/.

[36] 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, Heinrich Heine University Düsseldorf, 2021, pp. 69–75. URL: https://netlibrary.aau.at/obvukloa/download/pdf/6435190?originalFilename=true.