Exploring Intensities of Hate Speech on Social Media: A Case Study on Explaining Multilingual Models with XAI

Raisa Romanov Geleta,[1] **Klaus Eckelt,**[2] **Emilia Parada-Cabaleiro**[1,3,4] **and Markus Schedl**[1,3]

[1]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2]Institute of Computer Graphics, Johannes Kepler University Linz, Austria
[3]Human-centered AI Group, Linz Institute of Technology (LIT), Austria
[4]Department of Music Pedagogy, Nuremberg University of Music, Germany
raisa.geleta@gmail.com, klaus.eckelt@jku.at

Abstract

Hate speech on social media platforms has grown to become a major problem. In this study, we explore strategies to efficiently lessen its harmful effects by supporting content moderation through machine learning (ML). In order to present a more accurate spectrum of severity and surmount the constraints of seeing hate speech as a binary task (as typical in sentiment analysis), we classify hate speech into four intensities: no hate, intimidation, offense or discrimination, and promotion of violence. For this, we first involve 31 users in annotating a dataset in English and German. To promote interpretability and transparency, we integrate our ML system in a dashboard provided with explainable AI (XAI). By performing a case study with 40 non-experts moderators, we evaluated the efficacy of the proposed XAI dashboard in supporting content moderation. Our results suggest that assessing hate intensities is important for content moderators, as these can be related to specific penalties. Similarly, XAI seems to be a promising method to improve ML trustworthiness, by this, facilitating moderators' well-informed decision-making.

1 Introduction

The rapid growth of hate speech is a worrying problem that has been brought on by the immediate nature of social media (Mollas et al., 2022). Effectively limiting hate speech has become more difficult due to its wide impact and quick propagation (United Nation, 2023). Therefore, given the pressing need to address this issue, investigating efficient techniques and methodologies able to reduce its negative consequences has become crucial. By analyzing hate speech detection methods and the potential for XAI to improve transparency and interpretability, our study intends to support these initiatives.

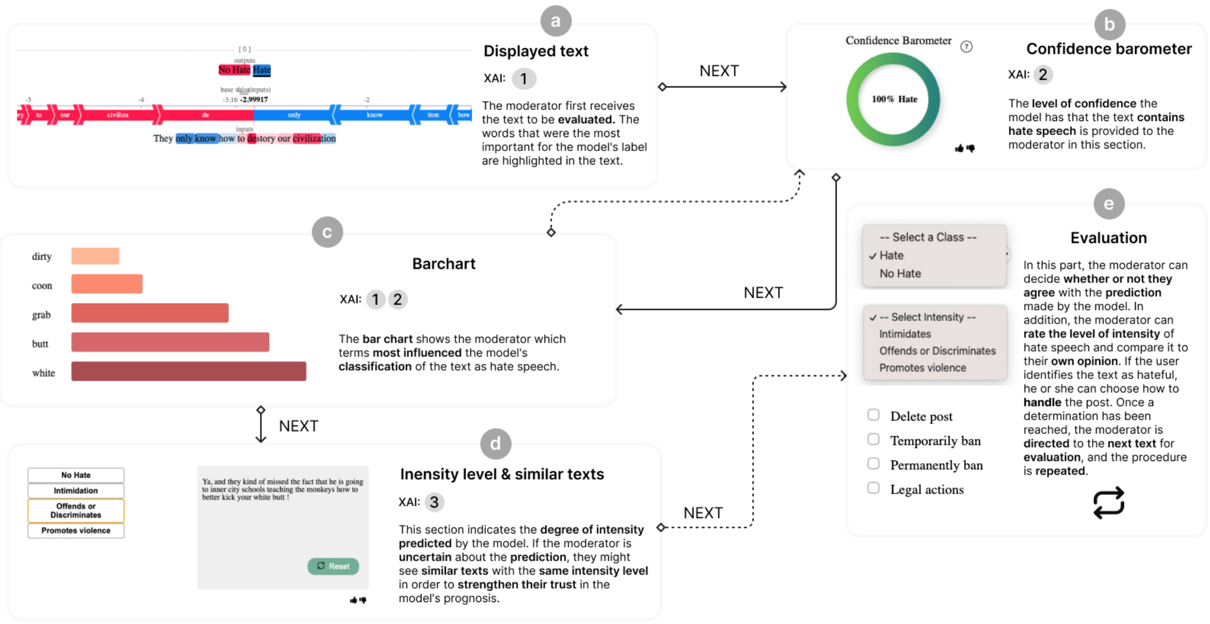
Hate speech is typically characterized in research studies as either being hateful or not, i. e., in binary terms (Aluru et al., 2021; Deshpande et al., 2022; Duwairi et al., 2021; Roy et al., 2020; Plaza-del Arco et al., 2021; Del Vigna et al., 2017). Nonetheless, there have been instances where more nuanced classifications have been examined (Ibrohim and Budi, 2019; Mollas et al., 2022; Del Vigna et al., 2017). To get over this limitation, we adopted the levels by Olteanu et al. (2018), which include three unique intensities: intimidation, offense or discrimination, and promotion of violence. In addition, we included "no hate" to account for situations in which hate speech traits are not present.

Through the design science research (DSR) methodology (Peffers et al., 2007), we create an artifact that engages humans in the evaluation of hate speech, i. e., a dashboard to support social media content moderation. Inspired by Bunde (2021), our dashboard (depicted in Figure 1) includes novel features, such as a hate speech detection algorithm based on Universal Language Model Fine-Tuning, SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017) text heat mapping, text similarity, and a four-level hate speech intensity scale. Our dashboard enables moderators to comprehend and explore the underlying assumptions of the machine learning (ML) model's predictions, by this assisting them in making well-informed decisions.

We aim to answer two Research Questions:
RQ1: Are intensities of hate speech an important factor to be considered in content moderation?
RQ2: Is XAI a successful way to support moderators' judgment of social media content?

2 Related Work

A well-defined, linguistically nuanced, and intergroup-relationship-aware concept is required for an automated approach to be precise (Fortuna



The moderator first receives the text to be **evaluated.** The words that were the most important for the model's label are highlighted in the text.

The **level of confidence** the model has that the text **contains hate speech** is provided to the moderator in this section.

The **bar chart** shows the moderator which terms **most influenced** the model's **classification** of the text as hate speech.

Ya, and they kind of missed the fact that he is going to inner city schools teaching the monkeys how to better kick your white butt !

This section indicates the **degree of intensity predicted** by the model. If the moderator is **uncertain** about the **prediction**, they might see **similar texts** with the **same intensity level** in order to **strengthen their trust** in the model's prognosis.

In this part, the moderator can decide **whether or not they agree** with the **prediction** made by the model. In addition, the moderator can **rate the level of intensity** of hate speech and compare it to their **own opinion**. If the user identifies the text as hateful, he or she can choose how to **handle** the post. Once a determination has been reached, the moderator is **directed** to the **next text** for **evaluation**, and the procedure is **repeated**.

Figure 1: Chart flow diagram of the moderator‘s journey through the XAI dashboard.

and Nunes, 2018). Amongst the number of definitions proposed in the literature, Nobata et al. (2016) identifies hate speech as speech that disparages and attacks a group based on characteristics like ethnicity, religion, gender, or sexual orientation. Fortuna and Nunes (2018) defines it as language that criticizes or disparages groups based on particular traits: depending on the linguistic style, it might provoke violence or hate. Despite the attempts, hate speech detection is still limited by the lack of a distinct and widely accepted definition.

Besides the conceptual problems of defining hate speech, technical difficulties in detecting it include differences in training datasets as well as biases in ML algorithms (MacAvaney et al., 2019). In addition, developing a uniform method to identify hate speech is further impaired by the different laws regarding the right to free speech from different nations (United Nation, 2023). Still, the urgency of effectively combating hate speech on social media has led to the development of a variety of ML techniques aiming to automatically identify it. One approach for transparent hate speech detection is Masked Rationale Prediction(MRP), introduced by Kim et al. (2022). MRP uses context-relevant tokens and unmasked rationales to anticipate masked human rationales in order to reduce bias and increase explainability. To detect hate speech on Twitter, Zhang et al. (2018) devised a C-GRU, which combines a CNN and a gated recurrent network (GRU), while Khan et al. (2022) introduced a deep learning model called BiCHAT that combines contextual word representation, deep CNN, BiLSTM, and hierarchical attention to successfully detect hate speech in Twitter.

Despite the promising outcomes, the application of ML in detecting hate speech presents still limitations. Nobata et al. (2016) emphasized that some forms of hate speech are not sufficiently investigated. Furthermore, it is well-known that ML models are affected by biases that negatively impact the decision-making process (Molnar, 2022). The lack of transparency of many ML models makes it more difficult to spot and correct such biases. Due to this, works like the one Mehta and Passi (2022) and Bunde (2021) have started looking at the possibility of using XAI to enhance the interpretability of hate speech recognition systems.

3 Methods

3.1 Dataset of Hate Speech

Since it has been shown that hate speech recognition through ML can be affected by the target language (Aluru et al., 2021), we investigate two languages in our study. In order to create a metacorpus of hate speech in English and German, we collected pre-existing hate speech datasets in both languages, which included GermEval[1] (Wiegand, 2019), hasoc-fire-2020[2] (Dowlagar and Mamidi, 2021), UCSM-DUE GHSR[3] (Ross et al., 2016) and those by Davidson et al. (2017) and de Gibert et al. (2018). From each language, a total of 1,500

[1] https://github.com/uds-lsv/GermEval-2018-Data

[2] https://github.com/suman101112/hasoc-fire-2020

[3] https://github.com/UCSM-DUE/IWG_hatespeech_public

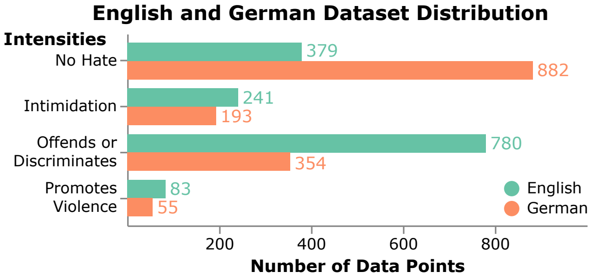


Figure 2: Distribution of annotator classifications for each intensity in English and German languages.

texts were randomly selected and annotated according to the labels proposed by Olteanu et al. (2018). Texts that contained only links or a username were removed, resulting in 1,437 and 1,476 samples for English and German, respectively. We reached out to potential annotators using social media sites including Instagram, Facebook, and Github. 31 contributors (18 males, 13 females, in a 26-35 age range) took part in the annotation process. A user interface was developed using Streamlit to enable users to annotate the data according to the hate intensity values. The application's source code is freely accessible.[4]

Before taking part in the experiment, the annotators were required to agree to the participation terms, which stipulated that their anonymous responses would be used for scientific research.[5] Each participant was instructed on the task before annotating a minimum of 10 samples in the chosen language. The annotators were requested to identify the level of hate expressed in the text through a forced-choice test. They could choose one of the following intensities: (i) *no hate*, (ii) *intimidation*, (iii) *offends or discriminates*, (iv) and *promotes violence*. The distribution of annotations across intensities and languages, shown in Figure 2, is highly imbalanced, which we expect to affect the ML performance. Compared to the other labels, the most extreme intensity *promotes violence* was chosen by far fewer times in both languages. The majority of German data was rated as *no hate*, whereas the majority of the English data was rated as *offends or discriminates*.

3.2 Dashboard

We developed an XAI dashboard[6] that supports multi-lingual evaluation to enhance content moderation strategies for safer online communities. Figure 1 depicts the interaction flow in the moderation dashboard. The first section (Fig. 1a) displays the input text, predicted label, and highlights the words that contributed to—or against—the prediction with a heatmap based on the words' SHAP values (Lundberg and Lee, 2017). We additionally calculate the predicted probabilities' entropy, with higher values indicating greater certainty, to assess the ML model's trustworthiness with the *Confidence barometer* (Fig. 1b) (Bogert, 2021). The bar chart in Figure 1c ranks the words most influential on the classification of *hate* or *no hate*. By visualizing the trustworthiness of the model and highlighting important words, users can make informed decisions and develop a deeper understanding of the underlying model.

The next section of the dashboard (Fig. 1d) displays the text's hate speech intensity and similar texts classified with the same intensity. A nearest neighbor search identifies text samples of similar content and hate intensity. These samples for the predicted intensity provide contextual information to enhance moderator precision.

The moderator can then evaluate the model's prediction and determine whether or not they concur with it (Fig. 1e). If the text is identified as non-hateful, the dashboard automatically directs the moderator to the next text. If the text is identified as hate speech, the moderator is prompted to select the level of hate speech intensity and decide on the appropriate action to take against the person who posted the text. The moderator can also rate the usefulness of the XAI methods and provide feedback by selecting the thumbs-up or thumbs-down icon next to each method (Fig. 1 1-4).

3.3 User Study

To test the XAI dashboard along with other evaluation methodologies we performed a user study with 40 volunteers (26 male, 14 female). Most of them were university students (n = 34) and around half Austrian (n = 22); the rest of participants were spread amongst 11 nationalities. Due to the imbalanced distribution, the potential effect of these attributes will not be evaluated. The individuals who exhibited the greatest level of skill in their particular languages were intentionally allocated to either the German or English cohort.

The goal of the user study was to assess whether different evaluation methodologies influence mod-

[4] https://github.com/Raisarom/Streamlit_AnnotationApp

[5] The procedures used in this research were carried out in accordance with the tenets of the Declaration of Helsinki.

[6] https://github.com/Raisarom/Hate-Speech-Detection-Dashboard-with-XAI

erators' decisions (see Figure 3). With evaluation methodologies, we refer to the underlying methods used to assign a hate label (suggested to the moderator) to a given text (presented to the moderator for evaluation). Four evaluation methodologies were assessed: A) labels suggested by a human; B) labels suggested by AI; C) labels suggested by a human who revised AI ratings; D) labels from AI assessed through the XAI dashboard. For each language, 10 participants were randomly assigned to each group. Their task was to act as "moderators" i. e., for a given text they would get a suggested label, and subsequently they were requested to rate the text. In case of disagreement w. r. t. the suggested label, they were requested to indicate the appropriate intensity of hate. To ensure an objective evaluation, moderators did not know to which group they were assigned.

3.4 ML Models Implementation

We implemented a system able to distinguish first between hate and no hate speech; subsequently between three fine-grained intensities (intimidates, offends, and promotes violence). Due to the limited size of our dataset, pre-trained Hate Bert Models from Huggingface were used to classify the data into hate and no hate, individually for each language (see Section 4). We also evaluated a multilingual Hate Bert model to test the machine's capacity to classify both languages together. The pre-trained models were fine-tuned with our re-annotated data of the respective language, or both languages for the multilingual model. The annotated data was also used to train several ML algorithms to additionally identify the hate intensity in the texts. These algorithms included Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Fasttext classifier, and a Dummy classifier used as baseline to evaluate the performance of the other classifiers. We opted for this two-step approach to leverage the information of the pre-trained models to improve the overall detection of hate speech and focussed on traditional algorithms instead of deep learning models due to the small size of the dataset and its imbalanced character.

Before training the models, the data was preprocessed following standard techniques in text processing, such as lowercase conversion, punctuation removal, stop-word removal, and lemmatization. The model performance will be evaluated in terms of precision, recall, F1 score, and accuracy metrics.

4 Results

4.1 ML Accuracy

In this study, separate BERT models were trained for each language to predict two output labels: hate and no hate. An approximate data split of 75-10-15 was aimed for, with slight deviations due to efforts to create a balanced test dataset. The distribution of sequence lengths in the dataset was examined to determine the optimal `max_length` for tokenization. The corresponding AutoTokenizer from the pre-trained BERT models was used, and the models were trained using `CrossEntropyLoss` and the `Adam optimizer`. Class weights were calculated based on the class distribution in the training set and added to the `CrossEntropyLoss` function to balance the contribution of each class during training. A `scheduler` was employed to adjust the learning rate during training. The training parameters provided by Liu et al. (2019) were followed.

In order to recognize the intensity of hate, we also trained a different model for each language. Due to space constraints only the optimal hyperparameters for the Random Forest classifier (which achieved best results) are given. According to the conducted grid search, the parameters were: max_depth $\in$ 20, min_samples_leaf $\in$ 2, min_samples_split $\in$ 2 and n_estimators $\in$ 100.

Table 1 shows the best performance by the pre-trained Hate BERT models for each language. While we also considered a model trained solely on English data,[7] the Multilingual-hatespeech-robacofi[8] Model (M-BERT) obtained the highest accuracy of 72% for the English dataset. The Bert-base-german-cased-hatespeechGermEval18Coarse2[9] Model (BERT-GER) achieved an accuracy of 68% in the German dataset. Overall, the Multilingual Bert Model outperformed the German one, especially in terms of precision and recall for the English data. Still, both models demonstrated comparable F1-scores. Among all the classifiers for hate speech intensity, the RF classifier achieved the highest accuracy with 38% on the English dataset and 48% on the German one. Note that in a three-class problem, these results, although low, are still above chance.

[7] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-english

[8] https://huggingface.co/Andrazp/multilingual-hate-speech-robacofi

[9] https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

Model Type	Dataset	Label	Precision	Recall	F1-Score	Accuracy
M-BERT	English	Hate	0.76	0.64	0.69	0.72
		No Hate	0.69	0.8	0.74	
BERT-GER	German	Hate	0.68	0.69	0.69	0.68
		No Hate	0.69	0.67	0.68	

Table 1: Performance of BERT models on English and German datasets for hate speech detection.

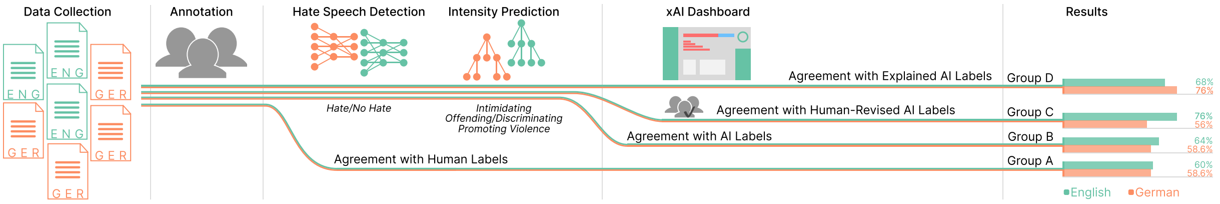


Figure 3: Design and results of the study comparing evaluation methodologies on the German and English datasets.

4.2 Dashboard Evaluation

We assessed the percentages of agreement within and across groups in order to evaluate each evaluation methodology's efficacy. The findings of our user case study are shown in Figure 3, along with the percentages of matches for each category and language. Groups A and B exhibited similar rates of agreement for the German group, however, Group C had a somewhat lower rate. With 76%, Group D had the highest level of agreement. The outcomes were a little different for the English group: Group A had the lowest match rate followed by Group B and Group D. The greatest match rate was in Group C with 76%.

Groups D and C had quite high agreement percentages. The results from Group D suggest that the dashboard's extra explanations enhance participants' confidence in their choices. Still, the results from Group C, highlight the importance of involving a person in the decision-making process.

Additionally, we looked into how the severity of hate speech related to moderator action. Spearman correlation indicated a smaller link between the intensity of hate speech and moderator actions in German ($r \approx 0.19$) than in English ($r \approx 0.54$).

5 Discussion and Limitations

The BERT model's inferior accuracy is probably due to the small amount of annotated data (about 1,450 data points), which constitutes one of the main limitations of our work. Indeed, larger datasets are often needed to attain the best performance for deep learning models like BERT, as shown in previous works (Saleh et al., 2023). Concerning the classification of hate intensity, the imbalance of our dataset contributed further to the low ML accuracy. There were remarkably few annotated data points, especially for the "promotes violence" category. Indeed, obtaining high-quality annotations for hate speech is a well-known problem, already highlighted by previous works (Del Vigna et al., 2017).

The outcomes from the user study revealed that there was a prominent bias toward political hate speech in the German data. This may, indeed restrict the usability of the German model in non-political hate speech, which highlights the need of collecting high-quality and representative dataset across multiple languages and contexts. Similarly, although the majority of study participants agreed with the utilized intensities, they also proposed adding others such as irony or sarcasm, which should be considered in the future research.

6 Conclusions

Concerning RQ1, our study shows that, especially for English, low hate intensities were generally related to moderator actions of low severity, such as *delete post* or *temporary ban*, while a higher hate intensity was mostly linked to permanent bans. This suggests that hate speech intensity might be a criteria to undertake specific moderator actions. Concerning RQ2, our results from the German data indicate that XAI improves the decision-making capabilities of moderators, as shown by a higher agreement with respect to the other methods.

We showed that defining hate speech in terms of intensities, as well as developing XAI tools, are both promising ways to improve the quality and effectiveness of online-content moderation, by this making the internet a safer place for everyone.

Acknowledgements

A special thanks to all of our participants. Without them, this study would have not been possible.

References

S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. 2021. A Deep Dive into Multilingual Hate Speech Classification. In *Proc. ECML PKDD*, pages 423–439.

K. Bogert. 2021. Notes on Generalizing the Maximum Entropy Principle to Uncertain Data. *arXiv*.

E. Bunde. 2021. AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach. In *Proc. HICSS*, pages 1264–1273.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proc. ICWSM*, pages 512–515.

O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proc. ALW2*, pages 11–20.

F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proc. ITASEC*, pages 86–91.

N. Deshpande, N. Farris, and V. Kumar. 2022. Highly Generalizable Models for Multilingual Hate Speech Detection. *arXiv*.

S. Dowlagar and R. Mamidi. 2021. HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection. *arXiv*.

R. Duwairi, A. Hayajneh, and M. Quwaider. 2021. A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. *Arab. J. Sci. Eng.*, 46:4001–4014.

P. Fortuna and S. Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4):1–30.

M. O. Ibrohim and I. Budi. 2019. Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter. In *Proc. ALW*, pages 46–57.

S. Khan, M. Fazil, V. Sejwal, M. Alshara, R. Alotaibi, A. Kamal, and A. Baig. 2022. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *J. King Saud Univ. Comput. Inf. Sci.*, 34:4335–4344.

J. Kim, B. Lee, and K. Sohn. 2022. Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection. In *Proc. COLING*, pages 6644–6655.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. NIPS*, pages 4765–4774.

S. MacAvaney, H. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. Hate Speech Detection: Challenges and Solutions. *PloS one*, 14(8):e0221152.

H. Mehta and K. Passi. 2022. Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). *Algorithms*, 15:291.

I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. 2022. ETHOS: A Multi-Label Hate Speech Detection Dataset. *Complex & Intell. Syst.*, pages 1–16.

C. Molnar. 2022. *Interpretable Machine Learning*, 2nd edition.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive Language Detection in Online User Content. In *Proc. WWW*, pages 145–153.

A. Olteanu, C. Castillo, J. Boy, and K. Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. In *Proc. ICWSM*.

K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, 24(3):45–77.

F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. Comparing Pre-Trained Language Models for Spanish Hate Speech Detection. *Expert Syst. Appl.*, 166:114120.

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proc. NLP4CMC*, volume 17, pages 6–9.

P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao. 2020. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. *IEEE Access*, 8:204951–204962.

H. Saleh, A. Alhothali, and K. Moria. 2023. Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model. *Appl. Artif. Intell.*, 37(1).

United Nation. 2023. UN Strategy and Plan of Action on Hate Speech. `https://www.un.org/en/hate-speech`.

M. Wiegand. 2019. GermEval-2018 Corpus (DE). heiDATA. V1.

Z. Zhang, D. Robinson, and J. Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-Gru Based Deep Neural Network. In *Proc. ESWC*, pages 745–760.