Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: "Are we on the same page ? "

Bruno Machado Carneiro, Michele Linardi, Julien Longhi

Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: *"Are we on the same page ? "*

Bruno Machado Carneiro, Michele Linardi, Julien Longhi
ENSEA Engineering School, ETIS UMR-8051 CY Cergy Paris Université, AGORA CY Cergy Paris Université
bruno.machadocarneiro@ensea.fr, {michele.linardi, julien.longhi}@cyu.fr

Abstract

We study Socially Unacceptable Discourse (SUD) characterization and detection in online text. We first build and present a novel corpus that contains a large variety of manually annotated texts from different online sources used so far in state-of-the-art Machine learning (ML) SUD detection solutions. This global context allows us to test the generalization ability of SUD classifiers that acquire knowledge around the same SUD categories, but from different contexts. From this perspective, we can analyze how (possibly) different annotation modalities influence SUD learning by discussing open challenges and open research directions. We also provide several data insights which can support domain experts in the annotation task. *Accepted for publication in the International Conference on CMC and Social Media Corpora for the Humanities (University of Mannheim, Germany, 2023).*

Keywords: SUD Classification, Machine Learning , Deep Learning, Transfer Learning, Annotation Guidelines

Acknowledgment

The work presented in this paper is part of the ARENAS project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No:101094731.

1. Introduction

During these last two decades, the massive popularisation of social media has been changing the way people communicate, interact and collect worldwide news. The dissemination speed rate and the possibility to quickly reach a large audience are some clear advantages of modern social network platforms. By contrast, the potential anonymity and sense of impunity can bring out the worst in people and made them sharing ideas that would not be socially acceptable otherwise. Socially Unacceptable Discourse (Sulc and de Maiti, 2020) (SUD) typically occur in various form; The use of offensive and abusive language represent a common form of SUD, but it is also important to note that controversial narratives are not necessarily bad or immoral, but they closely relate to radicalization and ideologies. Clear contexts in the recent history are the Covid-19 crisis and the the Russian invasion of Ukraine. During these periods, we have witnessed several cases of public debate radicalization, especially favored by the circulation of distorted information (De Giorgio et al., 2022) that jeopardizes the knowledge acquisition of complex systems and environments.

Another particular trait of SUD is the presence of distinctive grammatical characteristics. To model these features, we require identifying several grammatical substructures such as residual representations, use of pronouns, and future tense (Ascone and Longhi, 2018; de Maiti et al., 2020).

We note that, in publicly annotated corpus used so far by the Machine Learning community, no standard or common guidelines for SUD annotation exist (Fišer et al., 2017) despite the adoption of the same terminology and/or tags. It derives that different SUD definitions may potentially share overlapping characteristics, or on the other hand a single category may cover text instances with divergent features depending on the context. Furthermore, annotators bias can also play a decisive role as reported by previous works (Badjatiya et al., 2019; Yuan et al., 2022a; Davidson et al., 2019).

In this scenario, it is reasonable to expect a poor generalization capability of ML SUD classifiers trained in a specific context (Yuan and Rizoiu, 2022). To that extent, we study and evaluate the capability of current state-of-the-art Deep Learning models to characterize SUD on different grounds. Other works have recently considered the zero-shot learning problem in hate speech detection, where transfer learning is tested and measured on binary (hate/no hate) (Toraman et al., 2022) and on multi-class (Yuan and Rizoiu, 2022) classification. In this context, we sketch and propose a different approach that first aims to test transfer learning at a class level rather than a dataset level. This approach permits us to provide more interpretable insights on the SUD semantic and to test the transfer over different annotation guidelines on the same speech categories.

2. Socially Unacceptable Discourse Corpora

We report the corpora we consider in our study in table 1. We use data from various sources recently adopted to assess the performance of state-of-the-art ML solutions for automatic SUD detection (e.g., hate speech detection, sentiment, toxicity, radicalization, and ideology analysis).

We selected $\mathbf{13}$ publicly available datasets containing $\mathbf{470,768}$ samples distributed over 12 classes.

We generate a unique English text corpus by concatenating all the 13 datasets, denoting it with the label G^{SUD}. Note that the datasets we concatenate in G^{SUD} share multiple overlapping SUD labels, which identify the same SUD category. We consider the presence of bias and ambiguities as physiological, and identifying and analyse the concerned instances is under the lens of our research.

In figure 1(a), we report the instances distribution over SUD classes. Note that the *neither* class subsumes all texts that do not fall in any SUD categorizations proposed by the annotators. As expected, SUD classes have a sensitive lower support compared to the *neither* class denoting the typical class imbalance setting of the SUD detection problem.

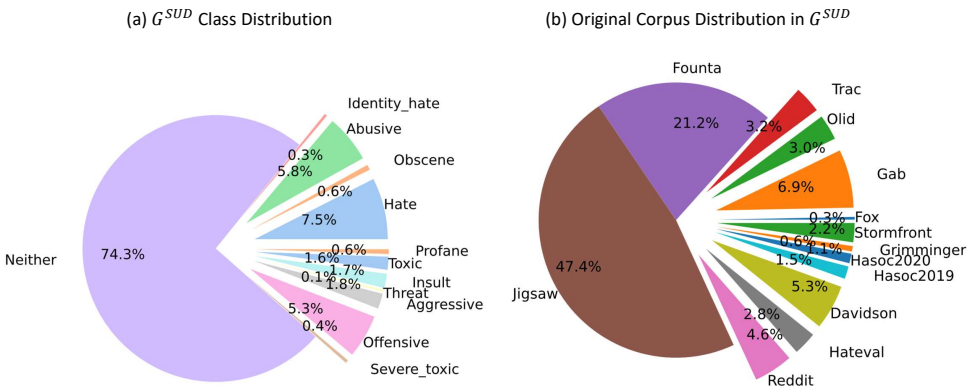

Figure 1: (a) G^{SUD} Class distribution, (b) Corpus distribution in G^{SUD}

Dataset	Sample type	# Samples	Topic	Best performing SUD classifier	F1 Macro (%)
Davidson (Grimminger and Klinger, 2021)	Tweets	25,000	Generic	BERT	93
Founta (Swamy et al., 2019)	Tweets	100,000	Generic	BERT	69.6
Fox (Yuan and Rizoiu, 2022)	Threads	1,528	Fox News Posts	BERT	65
Gab (Qian et al., 2019)	Posts	34,000	Generic	CNN	89.6
Grimminger (Grimminger and Klinger, 2021)	Tweets	3,000	US Presidential Election	BERT	74
HASOC2019 (Wang et al., 2019)	Facebook, Twitter posts	12,000	Generic	LSTM + Attention	78.8
HASOC2020 (Roy et al., 2021)	Facebook posts	12,000	Generic	XLM-RoBERTa	90.3
Hateval (MacAvaney et al., 2019)	Tweets	13,000	Misogynist and Racist content	mSVM/BERT	75.4
Jigsaw (van Aken et al., 2018)	Wikipedia talk pages	220,000	Generic	Bi-GRU + Attention	78.3
Olid (Zampieri et al., 2019)	Tweets	14,000	Generic	CNN	80
Reddit (Yuan and Rizoiu, 2022)	Posts	22000	Toxic subjects	BERT	85
Stormfront (MacAvaney et al., 2019)	Threads	10,500	White Supremacy Forum	BERT	80.3
Trac (Aroyehun and Gelbukh, 2018)	Facebook posts	15,000	Generic	LSTM	64

Table 1: Best performing SUD classification model on each dataset.

Figure 1(b) illustrates the ratio of each dataset with respect to the global corpus. We observe that Jigsaw and Founta contain together more than 60% of the data.

2.1. Datasets

Here, we provide the details of each dataset we join in G^{SUD}.

Davidson (Davidson et al., 2017) contains around 25,000 tweets labelled as being hateful, offensive or neither of those randomly sampled from a set of 85.4 million tweets produced by 33,458 different users. Each sample was labelled by at least three different annotators.

Founta (Founta et al., 2018) contains about 100,000 tweets, labeled with four categories: abusive, hateful, normal, and spam. In this dataset, a variable number of users (between five and ten) have annotated each sample.

Fox (Gao and Huang, 2018) contains 1528 comments posted on ten different popular threads on the Fox News website. In these data, two native English speakers have produced labels to differentiate hateful from normal content following the same annotation guidelines.

Gab (Qian et al., 2019) contains $34,000$ samples extracted form Gab, a social media, where users commonly share far-right ideologies (Jasser et al., 2021), annotated in the Amazon Mechanical Turk[1] platform, where at least 3 annotators provided a label for each sample.

Grimminger (Grimminger and Klinger, 2021) contains $3,000$ tweets on 2020 presidential election topic in the United States. The samples were labelled between hate speech or not by three undergraduate students, who discussed the annotation guidelines during the labelling process.

HASOC2019 (Modha et al., 2019) and **HASOC2020 (Mandl et al., 2020)** are datasets proposed in the Indo-European Languages (HASOC) challenge, which contain $12,000$ English text samples extracted from Twiter and Facebook labeled between hateful, offensive, profane or neither of those.

Hateval (Basile et al., 2019) gathers around $13,000$ tweets containing hateful and normal speech. The hateful content originates from accounts of potential victims of misogynism and racism.

Jigsaw[2] (van Aken et al., 2018) is a dataset provided in the Toxic Comment Classification Challenge. It contains about 220,000 samples extracted from Wikipedia talk pages differentiated into seven classes: toxic, severe toxic, obscene, threat, insult, identity hate, and neither of the previous.

Olid (Zampieri et al., 2019) contains $14,000$ tweets annotated using the Figure Eight Data Labelling platform [3]. In this context, tweet selection is executed by keyword filtering and human annotation.

Reddit (Qian et al., 2019) has 22,000 samples extracted from Reddit, labeled for hate speech detection by Amazon Mechanical Turk users. Before the labeling task, the text got selected according to a list of toxic subjects on the Reddit platform.

Stormfront (de Gibert et al., 2018) contains 10,500 samples taken from a white supremacy forum called Stormfront and divided into four classes: hate, no hate, related, and

[1] https://www.mturk.com/

[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[3] https://f8federal.com/

skip. The related class contains statements that can not be considered hateful without considering their context. Text belonging to the skip class does not contain enough information to determine if it can be classified as hateful.

Trac (Kumar et al., 2018) dataset gathers $15,000$ Facebook posts and comments classified into aggressive and non-aggressive language.

3. SUD Deep Learning Models

In this section, we introduce and describe the state-of-the-art Deep Learning models adopted for the SUD detection task in previous works. In Table 1, we show the best performer in each corpus. Here, we report the Macro F1 score, which is the recommended averaging method for F1 score when dealing with class imbalance. It is calculated by averaging the sum of the F1 score of each class.

Recall that the F1 score reports the harmonic mean of precision and recall of a classification model. For a particular input class, we compute the precision (P) and recall (R) of a SUD classifier as follows: $P = \frac{TP}{TP+FP}$, and $R = \frac{TP}{TP+FN}$, where TP denotes the number of correctly classified instances of the input class (true positive), FP denotes the number of occurrences that are wrongly assigned with the input class label (false positive), and FN represents the number the input class samples that are erroneously classified (false negative). Hence we have that $F1 = 2 \times \frac{P \times R}{P+R}$.

From Table 1, we observe that **BERT** (Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)) is the best performer model in the majority of the datasets. BERT adopts a Deep Learning (DL) architecture released by the Google AI Language team in early 2019, which is pre-trained by masked language model (MLM) and next sentence prediction (NSP) tasks over a large corpus of English data containing more than 3B words (Devlin et al., 2019). MLM consists of training the model to predict masked tokens in the corpus sentences, whereas the NSP training aims to predict if two sentences form a sequence in the original text. XLM-RoBERTa (Conneau et al., 2020) is a multilingual variant of the original BERT model.

BERT has clearly shown its superiority over other types of DL models previously adopted in SUD classification, such as Convolutional Neural Networks (CNN) (Qian et al., 2019) and Long-short term memory networks LSTM (Wang et al., 2019). The attention mechanism used by BERT represents a robust solution that can better learn long-range token dependencies, avoiding the limitation of LSTM networks, which assumes that each token depends only on previous ones. By contrast, BERT learns relationships considering all the tokens in a sentence simultaneously.

In this work, we evaluate the SUD classification performance of BERT in the heterogenous corpus we construct. In the next section, we present all the research questions we address, discussing the results we obtain.

4. Experiments

4.1. Multiclass SUD Classification

To conduct our experimental evaluation, we use the $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019; Yuan and Rizoiu, 2022) model pre-trained by WordPiece tokenizer algorithm. For the sake of reproducibility, we provide the code and the data used in the experiments along with the relative instructions in an online repository (Machado Carneiro et al., 2023).

	F1 Score (%)		
Training set	**Macro**	**Weighted**	**Micro**
G^{SUD}	53.9	86.8	87.1
G^{SUD} Balanced	51.3	85	84.5
G^{SUD} with Neither Undersampled	58.5	73.7	73.9
G^{SUD} balanced with Neither Undersampled	56.8	72.5	72.1
G^{SUD} (Binary classification)	88.5	91.3	91.2
G^{SUD} balanced (Binary classification)	89.7	89.7	89.7

Table 2: Comparison between all experiments

To perform SUD classification, we connect BERT pooled output layers to a Multi-Layer Perceptron (MLP) architecture that contains 12 output neurons (one per class). We have fine-tuned the MLP layer of proposed model on the G^{SUD} corpus using a $80\%/10\%/10\%$ splitting ratio for training, validation, and testing respectively. We have adopted a stratified sampling technique to keep the same class distribution throughout the three splits. Hyperparameters have been tuned by performing several complete training rounds, picking the setting with the best validation performance.

The research questions we want to address are the following: *Which are the state-of-the-art model generalization capability in a global context? What are the main challenges that hamper the SUD modelling effectiveness?*

Table 2 contains the results, where we report Macro, Weighted and Micro F1 score of the SUD classification. Note that the Weighted F1 weighs the global F1 average according each class support, whereas the Micro F1 score computes a global F1 making no distinction across classes. Considering that G^{SUD} contains highly unbalanced SUD classes, we repeat classification tasks after training our model on a balanced dataset. To that extent, we have performed random oversampling of minority classes as suggested by several works (Yuan and Rizoiu, 2022; Swamy et al., 2019; MacAvaney et al., 2019).

Furthermore, given the dominance of the *neither* class, we also consider a setting with under-sampled non-SUD text (neither class). Here, we have selected 10% of the non-SUD samples in a stratified way, maintaining the same proportion of the *neither* class samples in every dataset.

We note that undersampling the *neither* class has a sensitive effect on the model prediction capability as the Macro F1 score increases. On the other end, reducing the neutral class causes an increment of model errors for the *neither* class (majority class) as we observe a significant reduction of the Weighted and Micro F1 scores. It follows that coping with such an imbalance between non-SUD and SUD samples represents a concrete challenge (typically occurring in a real-world scenario), which is amplified in the extended corpus under consideration.

We also notice that producing a balanced class scenario by performing random oversampling does not provide any significant benefit. This suggests that class imbalance is only a joint cause of the model discrimination capability.

To better understand how the adopted model discriminates SUD classes, we visualize the generated text representa-

	Macro F1 Score (%)											
	Abusive	**Aggressive**	**Hate**	**Identity Hate**	**Insult**	**Neither**	**Obscene**	**Offensive**	**Profane**	**Severe Toxic**	**Threat**	**Toxic**
G^{SUD}	79.4	64.1	65.8	35.9	50	94.3	25.6	74.9	30.5	39.5	42.6	17.7
Davidson	-	-	41.4	-	-	88.5	-	89.2	-	-	-	-
Founta	81.7	-	33.2	-	-	95.5	-	-	-	-	-	-
Fox	-	-	13	-	-	82.6	-	-	-	-	-	-
Gab	-	-	86.4	-	-	88.6	-	-	-	-	-	-
Grimminger	-	-	10.8	-	-	93	-	-	-	-	-	-
HASOC2019	-	-	7.94	-	-	78.1	-	25	20.4	-	-	-
HASOC2020	-	-	6.67	-	-	91.1	-	29.7	39.1	-	-	-
Hateval	-	-	53.2	-	-	73.9	-	-	-	-	-	-
Jigsaw	-	-	-	37.9	53.1	97.5	26.9	-	-	40.4	46	18.1
Olid	-	-	-	-	-	85.8	-	45.3	-	-	-	-
Reddit	-	-	74	-	-	89.5	-	-	-	-	-	-
Stormfront	-	-	39.7	-	-	94.1	-	-	-	-	-	-
Trac	-	68.1	-	-	-	66.1	-	-	-	-	-	-

Table 3: Macro F1 Score of SUD classification per class and dataset.

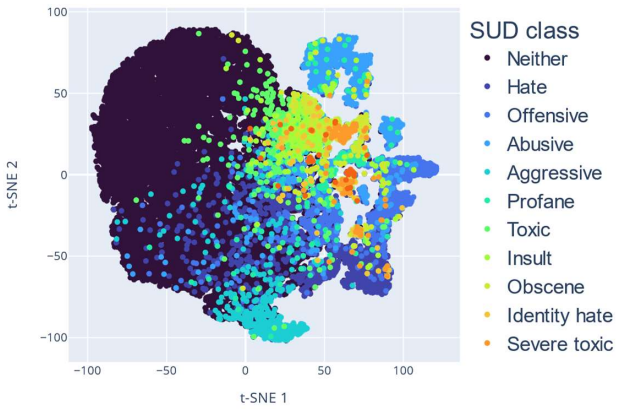

Figure 2: Two components t-SNE visualization of samples embedding produced by BERT output pooled layer.

tion (output of BERT output pooled layer). To reduce the dimensionality of the latent space, we apply t-distributed Stochastic Neighbor Embedding (t-SNE). Figure 2 shows the plot computed over the testing set, with a model trained on the complete corpus G^{SUD}. In Table 3 we report the Macro F1 score of SUD classification in G^{SUD} for each dataset and each class. Note that each line in this table corresponds to a different model, trained only on the specified dataset, while the first line is the result obtained using the model trained on G^{SUD}.

Here, we observe that some class features, i.e., *Abusive* (top-right), *Aggressive* (bottom-center) form fairly clear clusters. We can expect this behavior as each one of these class labels solely occurs in a single dataset, as depicted in Table 3.

Some other classes, i.e., *Hate*, *Offensive*, and *Toxic*, have more sparse values, which is one reason behind the absolutely low F1 score. Once again, these results get confirmed by the absolute low Macro F1 score both in the global corpus and in each single dataset.

Overall, the results explains the poor generalization capabilities of the studied classification model as this latter attains a low Macro F1 (58%) score on G^{SUD}. In detail, we note that problematic classes are not only those with the lowest number of training samples as one might expect. In fact, a performance drop occur in G^{SUD} classes that share samples from multiple corpus, suggesting the presence of intraclass heterogeneous samples as depicted in Table 3.

In this sense, a clear example concerns the *hate* class that contains samples from ten different datasets (out of thirteen). We note that shaky classification performance in each dataset of G^{SUD} (see Table 3) depends on divergent annotation criteria on a sensibly general concept, which can relate to different textual elements.

In Table 4, we depict the classification results obtained for each dataset in the global corpus G^{SUD}, and when the model was trained only using a single dataset (Individual). We note that only in two cases the global model performs better than the individual counterpart (for the Fox and Grimminger datasets). We believe that the relatively small support of these two corpora is the reason behind this improvement. Nevertheless, leveraging more knowledge from multiple domains does not constitute an advantage in practice.

	Macro F1 Score (%)			
	(a) Multiclass SUD Classification			**(b) Binary Classification**
Dataset	**Classified in G^{SUD}**	**Individual**		**Classified in G^{SUD}**
G^{SUD}	53.9	-		88.5
Davidson	73	75.1		93.9
Founta	70.1	74.7		92.9
Fox	**47.8**	41.6		59.2
Gab	87.5	89.9		86.2
Grimminger	**51.9**	46.9		64
HASOC2019	32.9	40.8		64.5
HASOC2020	41.7	48.4		88.2
Hateval	63.6	75.7		70.2
Jigsaw	45.7	52.6		87.7
Olid	65.6	75.2		72.3
Reddit	81.7	82.9		79.9
Stormfront	66.9	76.1		71.1
Trac	67.1	73.1		69.3

Table 4: (a) **Multiclass** SUD classification results (F1 score) with the model trained in G^{SUD} VS on each single dataset. (a) **Binary** SUD classification with the model trained in G^{SUD}.

4.2. Binary SUD Classification

For each of the experiments reported in this section, we have also tested the capability of the model to discriminate SUD and non-SUD text in G^{SUD} irrespective of the specific class. To that extent, we use the same configuration for the classification head, changing the output layer

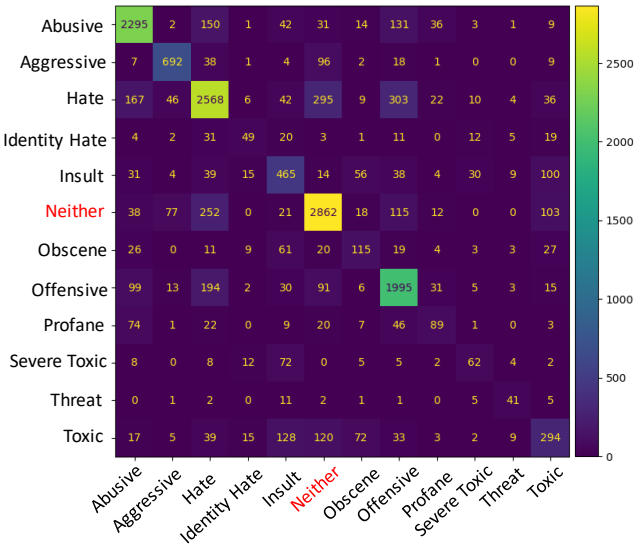

Figure 3: Confusion matrix of multi-class SUD classification.

to perform binary classification and re-training the model. For this case, we obtain a relatively high Macro F1 score ($\sim$ 90%). Such results suggest how the model discriminates well the *neither* class from the generic SUD in the global context we built, confirming the current trend observed in the ML literature so far. At the same time, effectively modeling multi-class SUD remains an open challenge.

5. Further Discussion and Perspective

To closely analyze the state-of-the-art limitation on SUD modeling, in figure 3, we plot the confusion matrix computed on the test set. In this case, we consider a test corpus with undersampled instances of neither class since, for this case, the classification model performs (slightly) in the best manner. Here, we can observe multiple critical cases that concern the labels *Identity Hate*, *Toxic*, *Obscene* and *Profane*. The classification model assigns a random label to these four classes that have overlapping features with all the others. Concerning classification performance, we note that the F1 score is not significantly dropping for these classes when the model applies to G^{SUD}. It derives that learned features are fairly conserved in the new global context.

This observation confirms the results proposed by prior studies (Yuan et al., 2022b; Fortuna et al., 2020), which already analyzed the relation among several classes in significantly smaller corpora.

We believe the large-scale scenario we propose motivates the need for a more consistent effort in the ML community to equip language models with more discriminant power. This concerns the capability to distinguish the source and the target of the SUD discourse (individual rather than group), as well as the elements that characterize the kind of narrative of each SUD class.

6. Conclusion and Future Work

In this work, we present an empirical evaluation of automatic SUD detection using the BERT model, a state-of-the-art Deep Learning architecture for SUD classification. To test generalization capability, we consider a large and heterogeneous context in which we obtain results that are not in line with the expected performance of the model trained at the local level, i.e., in every single corpus. In this sense, we argue that to build more general and reliable models, the ML community should consider formal guidelines provided by language experts (mostly neglected so far), which can sensibly reduce local bias (e.g., annotation policy, context, etc.). In future work, we plan to closely analyze the inter-domain mismatches we observe at the class sample level. Such effort would be beneficial to understand how to improve textual feature learning and to communicate requirements and expectations from the annotation task.

We furthermore note that the results and the insights we obtained also have the potential for the research linguists, discourse analysis, or semantics, as they show, from a knowledge base constituted by the main works on SUD corpora, the semantic links, and conceptual relationships, between several labels or tags.

In fact, over and above terminology, it is crucial to clearly state and understand the specific features of hate speech, offensive speech, or extremist speech. These initial results are necessary to foster several research discussions in the Horizon Europe ARENAS project into which this work integrates.

Specifically, the semantic issues in discourse categorization have an impact not only in terminological and computational terms (for annotating and classifying) but also in legal, political, and sociological terms. The impact of different characterizations is not neutral, there are potential issues of moderation or condemnation (Longhi, 2021), and it is necessary to proceed cautiously and rigorously in the delimitation of the chosen descriptors and in the way they are defined and characterized.

Finally, the explicability of these categories and the classification provided by Artificial Intelligence is central to future research. Making transparent outcomes will enable us to propose valuable results for all those involved in the hate speech and extremism analysis. In the context of a multidisciplinary project like ARENAS, which brings together scientists with different backgrounds (i.e., linguists, political scientists, etc.) and targets a heterogeneous audience, such as lawyers and journalists, the clarity of descriptors, and their ability to be understood by different stakeholders, is an essential element.

7. References

Aroyehun, S. T. and Gelbukh, A. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *TRAC-2018*.

Ascone, L. and Longhi, J. (2018). The expression of threat in jihadist propaganda. *Fragmentum*.

Badjatiya, P., Gupta, M., and Varma, V. (2019). Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter.

In *Proceedings of the 13th International Workshop on Semantic Evaluation*, June.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL 2020, Online, July 5-10, 2020*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *ALW2*, October.

De Giorgio, A., Kuvašić, G., Maleš, D., Vecchio, I., Tornali, C., Ishac, W., Ramaci, T., Barattucci, M., and Milavić, B. (2022). Willingness to receive covid-19 booster vaccine: Associations between green-pass, social media information, anti-vax beliefs, and emotional balance. *Vaccines*, 10.

de Maiti, K. P., Fišer, D., and Erjavec, T. (2020). Grammatical footprint of socially unacceptable facebook comments. In *Language Technologies & Digital Humanities*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*.

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *International Conference on Language Resources and Evaluation*.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior.

Gao, L. and Huang, R. (2018). Detecting online hate speech using context aware models.

Grimminger, L. and Klinger, R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, April.

Jasser, G., McSwiney, J., Pertwee, E., and Zannettou, S. (2021). Welcome to #gabfam: Far-right virtual community on gab. *New Media & Society*.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *(LREC 2018)*.

Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, 318:110564.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*.

Machado Carneiro, B., Linardi, M., and Longhi, J. (2023). `https://github.com/mlinardiCYU/SUD_study_different_eyes.git`.

Mandl, T., Modhab, S., Shahic, G. K., Jaiswald, A. K., Nandinie, D., Patelf, D., Majumderg, P., and Schäfera, J. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages.

Modha, S., Mandl, T., Majumder, P., and Pate, D. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages.

Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech.

Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021). Leveraging multilingual transformers for hate speech detection.

Sulc, A. and de Maiti, K. P. (2020). No room for hate: What research about hate speech taught us about collaboration? In *TwinTalks@DH/DHN*.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Toraman, C., Şahinuç, F., and Yilmaz, E. H. (2022). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference*.

van Aken, B., Risch, J., Krestel, R., and Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis.

Wang, B., Ding, Y., Liu, S., and Zhou, X. (2019). Ynu_wb at hasoc 2019: Ordered neurons lstm with attention for identifying hate speech and offensive language. In *Fire*.

Yuan, L. and Rizoiu, M.-A. (2022). Detect hate speech in unseen domains using multi-task learning: A case study of political public figures.

Yuan, L., Wang, T., Ferraro, G., Suominen, H., and Rizoiu, M.-A. (2022a). Transfer learning for hate speech detection in social media.

Yuan, S., Maronikolakis, A., and Schütze, H. (2022b). Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. 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)*.