Dimitra Niaouri, Bruno Machado Carneiro, Michele Linardi,
and Julien Longhi

Machine Learning is heading to the SUD (Socially Unacceptable Discourse) analysis: From Shallow Learning to Large Language Models to the rescue, where do we stand?

Abstract: The rapid proliferation of social media platforms has led to a significant increase in online Socially Unacceptable Discourse (SUD). SUD, characterized by offensive language, controversial narratives, and distinct grammatical patterns, poses a substantial challenge for online platforms. Effective detection of SUD necessitates robust Machine Learning (ML) models capable of generalizing across diverse contexts and performing well in binary and multi-class classification. The absence of standardized annotation guidelines and the variability of annotation modalities in existing corpora impede the development of such models in a large-scale scenario typically found in multiple online scenarios (e.g., social media platforms).

This research introduces a comprehensive corpus of manually annotated texts from various online sources to facilitate a thorough benchmarking of state-of-the-art SUD classifiers across twelve distinct discourse categories. We provide a novel comparative analysis of three model families: Shallow Learning Models (SLMs), Masked Language Models (MLMs), and Causal Language Models (CLMs), including models such as Support Vector Machines (SVM), Multinomial Logistic Regression (MLR), BERT, and BERT variants (ALBERT, RoBERTa, ELECTRA), Llama 2, and Mistral among others. We assess the performance of these models in binary and multi-class large classification scenarios, moving beyond the standard binary and limited-class frameworks existing in the literature. We further extend our analyses through various experimental scenarios, including the impact of class imbalances, and enhance model explainability. By applying visualization techniques to the text representations generated by the top-performing model, we observe class overlap and evaluate the model's generalizability.

Dimitra Niaouri, AGORA, CY Cergy Paris Université, e-mail: dimitra.niaouri@cyu.fr ETIS UMR-8051
Bruno Machado Carneiro, ENSEA Engineering School, e-mail: bruno.machadocarneiro@ensea.fr
Michele Linardi, CY Cergy Paris Université, e-mail: michele.linardi@cyu.fr ETIS UMR-8051
Julien Longhi, AGORA, CY Cergy Paris Université, e-mail: julien.longhi@cyu.fr

https://doi.org/10.1515/9783111434018-010

Our findings reveal limitations in current Deep Learning (DL) models for SUD classification due to class imbalances and inconsistent annotation guidelines. While binary SUD classification demonstrates promise, sensitivity to class imbalance in multi-class scenarios underscores the need for improved discriminatory power. Our analysis highlights the trade-off between bidirectional contextual awareness (favoring MLMs) and sequential dependency modeling (advantageous for CLMs), with MLMs emerging as the superior choice due to their bidirectional training approach. Finally, we emphasize the importance of consistent efforts within the ML community and the broader implications for linguistics, discourse analysis, and semantics, advocating for developing formal guidelines.

Keywords: Socially Unacceptable Discourse Analysis, Machine Learning, Deep Learning, Multi-source learning, corpus, Masked Language Models, Causal Language Models

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 make them share ideas that would not be socially acceptable otherwise. As a result, accurate detection and characterization of harmful ideas is crucial for effective social media moderation (Badjatiya et al. 2017; MacAvaney et al. 2019; Röttger et al. 2021; Alkomah and Ma 2022) as it enables targeted interventions, uncovers underlying issues such as prejudice, and supports the development of legal frameworks.

Although Machine Learning (ML) shows potential for automating content detection, there are substantial challenges that limit its effectiveness. Analysts encounter numerous overarching issues when using current ML solutions to detect Socially Unacceptable Discourse (Sulc and Pahor De Maiti 2020) (SUD), which often manifests in different forms and data modalities (Gandhi et al. 2024). A common form of SUD is the use of offensive and abusive language. However, it is important to note that controversial narratives, while not inherently bad or immoral, often have a close connection to radicalization and extremist ideologies. This relationship has become particularly evident in recent historical contexts such as the Covid-19 crisis and the Russian invasion of Ukraine, during which we have witnessed several cases

of public debate radicalization, especially favored by the circulation of distorted information (De Giorgio et al. 2022). Another particular trait of SUD is the presence of distinctive grammatical characteristics. To accurately model these features, it is essential to identify specific grammatical substructures, including residual representations, pronoun usage, and future tense (Ascone and Longhi 2018; Pahor De Maiti et al. 2020). Despite this, current publicly annotated corpora used in ML lack standardized guidelines for SUD annotation (Fišer, Erjavec and Ljubešić 2017). While similar terminology or tags are employed, different definitions of SUD may share overlapping characteristics, or a single category may encompass text instances with divergent features depending on the context. Moreover, annotator bias, as highlighted in previous studies (Badjatiya, Gupta and Varma 2019; Yuan et al. 2023; Davidson, Bhattacharya and Weber 2019), can significantly affect the consistency and accuracy of SUD annotations. As Yu et al. 2024 suggest, the primary data quality issues impacting model performance are noisy annotations, class imbalance, and data homogeneity.

Other complex forms of socially unacceptable discourse have recently started to receive attention. One example is the concept of extremist narrative, which identifies online discourses related to multiple social processes like radicalization, populism, demagogy, and other manifestations that endanger democracy. The ARENAS European project aims to significantly advance the extremist narrative analysis (Postigo-Fuentes et al. 2024). One of the main objectives consists of developing strategies for identifying, analyzing, and countering extremist rhetoric, seeking to advance beyond traditional methods by exploring the complex relationship between language and ideology in extremist content.

In this novel context, it is crucial to propose and assess ML solutions that support practical strategies for the accurate classification of multiple kinds of discourse, whose characterization depends on the social phenomenon, political scenario, and legal framework but also on the context, speaker, and intent of the speech itself. 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 (SOTA) ML models to characterize SUD within a large-scale, multi-class framework that better reflects real-world scenarios, where naturally multiple distributions exist. The rationale behind this approach is that the diverse range of discourse and topics in such a framework pose challenges to models to adapt, highlighting limitations in automatic detection and paving the way for improvements. Such effort will permit us to define research directions and open challenges to better address imminent requirements in SUD and extremist narrative analysis.

Given the limited availability of high-quality data for SUD detection, we note that transfer learning provides a well-established solution that leverages models

trained on datasets from related domains. Such an approach significantly reduces the requirement for extensive labeled examples in the specific target domain (Neyshabur, Sedghi and Zhang 2020).

In our evaluation, rather than considering cross-domain transfer learning, which consists of training a model on a single distribution and testing on another one from a different domain (e.g., Karan and Šnajder 2018; Swamy, Jamatia and Gambäck 2019), we implement a methodology that evaluates the capacity of SOTA models to generalize to SUD classes that naturally occur in multiple distributions (different contexts). Beyond the standard evaluation of models through intra-dataset classification, we also use an inter-dataset classification method. Our approach considers a large dataset resulting from the union of multiple datasets encompassing 12 classes. This methodology allows us to propose interpretable insights into the semantics of SUD and enables the evaluation of pattern learning across different annotation guidelines.

In the intra-dataset classification task, we maintain the same data split (training/test/validation) used in the performance assessment of each single dataset.

Our contributions can be summarized as follows:

1. We construct a unified corpus (G^{SUD}) from 13 publicly available datasets to fine-tune and evaluate pre-trained LLMs for general tasks and Shallow learning models at an intra- and inter-dataset level where the main focus is the generalization over classes rather than datasets.
2. We perform an extensive empirical evaluation of 12 SOTA models in the large-scale (~500K samples) multi-class scenario in G^{SUD}, moving beyond the standard binary and limited-class frameworks existing in the literature that typically involve fewer samples and narrower coverage of the intricate aspects of SUD.
3. We provide a unique comparative analysis of three model families: Shallow Learning Models (SLMs), Masked Language Models (MLMs), and Causal Language Models (CLMs) under a wide range of experimental conditions, including class imbalances, after tweaking the G^{SUD} dataset.
4. We enhance model explainability by employing visualization techniques on the text representations generated by the best-performing model, allowing us to observe class overlap and assess the model's generalizability over different SUD categories.

2 Related Work

Most prior research in cross-dataset and cross-domain generalization has focused on evaluating models trained on one dataset and tested on another, often within binary or limited-class scenarios.

Gröndahl et al. (2018) have explored cross-dataset generalization by replicating seven Machine Learning (ML) and Deep Learning (DL) models, including Logistic Regression (LR), Multi-layer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). The proposed benchmark considers four datasets from Wikipedia and Twitter in a binary classification setup. The study concludes that transferring knowledge between datasets results in poorer performance than training and testing on the same dataset. Additionally, simpler architectures performed comparably to more complex ones.

Similarly, Karan and Šnajder (2018) investigated generalizability across nine different datasets, including sources such as newspapers, Fox News, Twitter, and Wikipedia, comparing various SVM classifiers under a binary class setup. Their study reaffirms the challenges of cross-domain generalization, noting that models consistently performed better on in-domain test sets than on out-domain ones.

Swamy, Jamatia and Gambäck (2019) also contributed to the discourse by examining cross-dataset generalization using several ML and DL models, including LR, RNN, LSTM, ELMo, and BERT, across four different datasets in a positive vs negative class configuration. Their findings aligned with previous studies, highlighting that BERT was the best-performing model. They also observed that datasets with more positive samples generalized better and noted a significant drop in performance when transitioning from large training datasets to smaller test sets.

Pamungkas and Patti (2019) extended the exploration of generalization to cross-domain and cross-linguistic contexts, using ten publicly available datasets, which they binarized. They hypothesized that training on a dataset with broader coverage and testing on a narrower one would yield better results. Despite the out-domain scenario leading to worse performance, this work demonstrates that broader datasets enhance generalization compared to narrower ones. Salminen et al. (2020) further addressed the lack of cross-platform model development and testing, creating a cross-platform online hate classifier. They employed several ML algorithms (e.g., LR, Naive Bayes (NB), Support Vector Machine (SVM), XGBoost, Feed-Forward Neural Network (FFNN)) in a binary setup using an aggregated dataset from multiple platforms. XGBoost outperformed other models, with BERT-based features yielding the best results. Markov and Daelemans (2021) focused on reducing false positives in hate speech detection, evaluating various ML and DL models, including Bag-of-Words (BOW), CNN, LSTM, SVM, BERT, and RoBERTa. Under a binary setup, BERT and RoBERTa outperformed baselines and SVM in in-domain

conditions, though performance dropped in out-domain conditions, with BERT and RoBERTa still leading.

Similarly, to previous studies, Fortuna, Soler-Company and Wanner (2021) conducted experiments with several ML and DL models, including BERT, ALBERT, SVM, and fastText, standardizing dataset labels for intra- and cross-dataset setups. They advanced beyond typical single-dataset studies by examining nine datasets, focusing on those providing the highest generalization. Their results suggested that BERT and ALBERT outperformed the other two models under an intra-dataset classification scenario, while at the same time, they generalize better under an inter-dataset classification one. Yin and Zubiaga (2021) have focused on discovering the factors constraining model generalizability across datasets, highlighting challenges such as differing topics, label definitions, and data source platforms. They found that broader labels facilitated higher generalizability, with models like BERT and ALBERT performing relatively well. Toraman, Şahinuç and Yılmaz (2022) moved towards more complex classifications by creating large-scale tweet datasets in English and Turkish, covering five domains, to analyze the performance of various models for hate speech detection. They used a three-class setup (hate, offensive, normal) and evaluated traditional algorithms (LR), neural networks (CNN, LSTM), and transformers (BERT). As expected, the Transformer architecture outperformed simpler models, although the latter remained competitive.

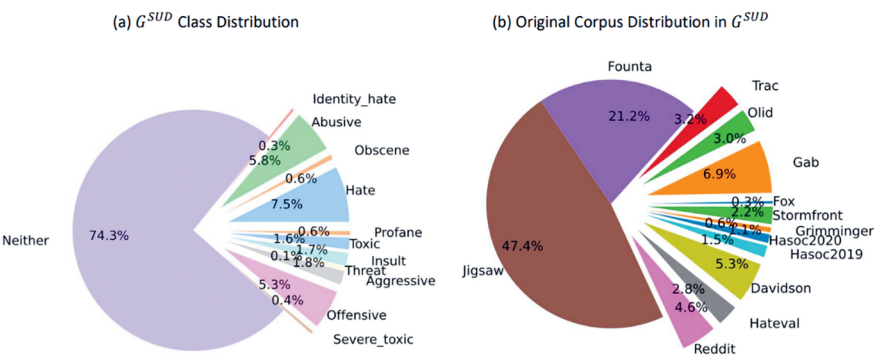

Figure 1: (a) G^{SUD} Class distribution, (b) Corpus distribution in G^{SUD}.

Antypas and Camacho-Collados (2023) took a significant step by examining generalizability across 13 hate speech-related social media datasets, using binary (hate vs. not hate) and multi-class (seven classes including racism, sexism, etc.) settings. They fine-tuned SVM, BERT, RoBERTa, TimeLMs-21, and BERTweet models on individual and unified datasets. Their findings showed that, under the binary and multi-class

setup, Transformer models achieve higher performances when trained on the combined dataset rather than on individual test sets different from their training sets. Gandhi et al. (2024) have recently made a step towards exploring multi-class classification. In their study, they consider a dataset of approximately 20,000 samples encompassing various classes, including hate speech, abusive language, individual and group hate, religious hate, and race-based hate, among others. Despite the broad range of classes, the benchmark tested a limited number of models, namely logistic regression and LSTM. Additionally, the research did not address cross-dataset and cross-domain generalization challenges. Finally, Yigezu et al. 2023 focused on multi-class and multi-label hate speech detection against the Mexican Spanish-speaking LGBTQ+ population using BERT and RoBERTa models. They used three classes (LGBTQ+ phobic, not LGBTQ+ phobic, NA) for multi-class and distinguished between various phobias (e.g., lesbophobia, gayphobia) for multi-label. BERT excelled in multi-label tasks, while RoBERTa was superior for multi-class tasks.

In contrast to these studies, our work overcomes previous research limitations by focusing on a multi-class scenario using a consistently larger scenario than the ones considered until now. We employ a unified dataset to develop general models subsequently tested on individual and smaller test sets. Such a choice enables us to capture the real-world complexities of SUD detection and to understand model generalization limitations in a multi-class context, offering novel insights compared to previous studies. In the following sections, we will delve into the datasets utilized and the specificities of our methodology, providing a detailed account of how our approach advances the field.

3 SUD corpora

Many works have proposed annotated datasets for hate speech analysis (e.g., Davidson et al. 2017; Founta et al. 2018; Qian et al. 2019; Grimminger and Klinger 2021). Among the most recognized resources is "hatespeechdata",[1] which compiles various dataset publications and their links. Poletto et al. (2020) conducted a comprehensive survey of available corpora, highlighting key benchmark datasets for evaluating abusive language. More recently, Piot, Martín-Rodilla and Parapar (2024) compiled an updated collection of over 60 datasets, named MetaHate, focusing on detecting harmful online content, including hate speech and cyberbullying, and analyzing text across social media platforms.

1 https://hatespeechdata.com/ (last accessed 14 February 2025).

In Table 1, we report the corpora we consider in our study. We use data from various sources recently adopted to assess the performance of SOTA ML solutions for SUD detection (e.g., hate speech detection, sentiment, toxicity, radicalization, and ideology analysis). We selected 13 publicly available datasets containing 470,768 samples distributed over 12 classes to advance beyond the binary classifications and limited class scopes of earlier research which generally involve fewer samples and less comprehensive coverage of hate speech scenarios. Our dataset choices are based on their comprehensive coverage of various aspects of SUD and their availability in English. By concatenating these 13 datasets, we create a unique English text corpus, which we have labeled 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 analyzing 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.

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 85.40 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[2] platform, where at least 3 annotators provided a label for each sample.

2 https://www.mturk.com/ (last accessed 14 February 2025).

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

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	69.60
Fox (Yuan and Rizoiu 2022)	Threads	1,528	Fox News Posts	BERT	65
Gab (Qian et al. 2019)	Posts	34,000	Generic	CNN	89.60
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	78.80
HASOC2020 (Roy et al. 2021)	Facebook posts	12,000	Generic	XLM-RoBERTa	90.30
Hateval (MacAvaney et al. 2019)	Tweets	13,000	Misogynist and Racist content	mSVM/BERT	75.40
Jigsaw (van Aken et al. 2018)	Wikipedia talk pages	220,000	Generic	Bi-GRU + Attention	78.30
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	80.30
Trac (Aroyehun and Gelbukh 2018)	Facebook posts	15,000	Generic	LSTM	64

Grimminger (Grimminger and Klinger 2021) contains 3,000 tweets in 2020 presidential election topic in the United States. The samples were labelled as 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 Twitter and Facebook labeled as 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[3] (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.[4] 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 skip. The related class contains statements that cannot 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.

4 SUD models

In this section, we examine SOTA models used for SUD detection. In section 3.1, we present the SOTA DL models adopted for the SUD detection task in previous works, and in section 3.2, we introduce the models we fine-tuned for this paper.

4.1 SOTA Deep Learning models

In Table 1, we show the best performer in each corpus. Here, we report the Macro F1 score used to evaluate the performance of our models. 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) of a SUD classifier as follows: $P = \frac{TP}{TP + FP}$ and recall (R) as: $R = \frac{TP}{TP + FN}$, where TP denotes the number of correctly classified

3 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (accessed 14 March 2025).
4 https://f8federal.com/ (last accessed 14 February 2025).

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 of 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 performing model in the majority of the datasets. BERT adopts a 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 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.

Table 2: Overview of the fine-tuned models.

Category	Models	Citation
SLM (Shallow Learning Models)	Gradient Boosting (GB)	Friedman (2001)
SLM (Shallow Learning Models)	Multinomial Logistic Regression (MLR)	Wright (1995)
SLM (Shallow Learning Models)	Multinomial Naive Bayes (MNB)	Kibriya et al. (2004)
SLM (Shallow Learning Models)	Random Forest (RF)	Breiman (2001)
SLM (Shallow Learning Models)	Support Vector Machines (SVM)	Hearst et al. (1998)
MLM (Masked Language Models)	$BERT_{BASE}$	Devlin et al. (2019)
MLM (Masked Language Models)	$ALBERT_{BASE}$	Lan et al. (2019)
MLM (Masked Language Models)	$RoBERTa_{BASE}$	Liu et al. (2019)
MLM (Masked Language Models)	$ELECTRA_{BASE}$	Clark et al. (2020)
CLM (Causal Language Models)	Llama-2-7b	Touvron et al. (2023)
CLM (Causal Language Models)	Mistral-7B-v0.1	Jiang et al. (2023)
CLM (Causal Language Models)	mpt-7b	MosaicML NLP Team (2023)

4.2 Fine-tuned models

In our study, we empirically evaluate the SUD classification performance of three different model families: Shallow Learning Models (SLMs), Masked Language Models (MLMs), and Causal Language Models (CLMs). An illustrative summary of the fine-tuned models can be found in Table 2. In the next sections, we elaborate on the specific characteristics of each category.

4.2.1 Shallow Learning Models

Shallow learning models represent a category encompassing conventional ML algorithms proposed prior to 2006 (Xu et al. 2021). This involves simple models with a few layers or processing units. They are suitable for tasks with straightforward data patterns, but their simplicity may limit their ability to capture complex relationships and adapt to new data. Hence, the performances of such models are closely tied to the effectiveness of the feature extraction process (Janiesch, Zschech and Heinrich). Within this overarching categorization, we specifically investigate Gradient Boosting (GB), Multinomial Logistic Regression (MLR), Multinomial Naive Bayes (MNB), Random Forest (RF), and Support Vector Machines (SVM).

4.2.2 Masked Language Models

Masked Language Models (MLMs), as explained in (Devlin et al. 2019), are DL models trained to reconstruct the original words of masked tokens based on the surrounding context. The significant advantage of those models lies in their bidirectional context, considering both preceding and subsequent tokens during the prediction process.

Within this category we evaluate $\text{BERT}_{\text{BASE}}$ (Devlin et al. 2019; Yuan and Rizoiu 2022) and some of the architectural variants introduced to enhance overall performance and reduce computational complexity. The BERT variants considered are the following:

1. $\text{ALBERT}_{\text{BASE}}$, which implements two parameter reduction techniques, namely Cross-layer parameter sharing and Factorized Embedding Parameterization. This results in a significantly smaller model compared to BERT (Lan et al. 2019). Moreover, ALBERT diverges from BERT's training approach by incorporating Sentence-Order Prediction (SOP) instead of Next Sentence Prediction (NSP).

2. RoBERTa$_{BASE}$, an optimized BERT pre-training approach, introduces several key modifications, including dynamic token masking for varying epochs, larger byte-level Byte-Pair Encoding (BPE), elimination of the NSP task, and an expanded corpus with increased training steps (Liu et al. 2019).
3. ELECTRA$_{BASE}$, another BERT variant, replaces the MLM training task with a replaced token detection task. This approach introduces binary classification to distinguish between original and replaced tokens, while omitting the NSP, aligning with the trends observed in ALBERT and RoBERTa (Clark et al. 2020). Despite its unique architecture that includes a replaced token detection task instead of a MLM one, ELECTRA is still categorized within our framework under the MLM family for ease of classification given its shared characteristics with BERT and other variants.

4.2.3 Causal Language Models

As introduced, MLMs are bidirectional models trained to consider context from both directions. Conversely, CLMs are unidirectional, considering only the previous context when making predictions. Specifically, they are trained to predict the next token in a sequence based on previous tokens, making them particularly efficient in text generation tasks. The CLM models fine-tuned and evaluated in this study are:

4. Llama 2 is a series of generative text models with varying parameters, ranging from 7 billion to 70 billion. Developed by Meta on an optimized transformer architecture (Touvron et al. 2023), these models are pre-trained and fine-tuned for language generation tasks. Notable innovations include pre-normalization using Root Mean Square Layer Normalization (RMS Norm), the use of Swigglue activation function, self-attention with KV Cache, and Rotary Positional Embedding (ROPE). For this study, we consider Llama 2 of size 7B parameter (Llama-2-7b-hf) fine-tuned on our datasets.
5. Mistral is a series of pre-trained generative text models developed by the Mistral AI team (Jiang et al. 2023). The model's innovation compared to Llama 2, as summarized in their paper, lies in its use of Sliding Window Attention (SWA), Rolling Buffer Cache, and Pre-fill and Chunking. Again, the size of the model is of 7B parameters (Mistral-7B-v0.1).
6. MPT, proposed by the MosaicML NLP team (2023) and released in various sizes and fine-tuned variations, constitutes another series of Large Language Models (LLMs). Adopting a GPT-style architecture with a decoder-only transformer, MPT features refinements such as performance-optimized layer implementations and architectural modifications for enhanced training stability among others. The 7B parameter sized model (mpt-7b) is tested in this study.

Table 3: Optimal performing SLM per class and dataset.

	Macro F1 Score												
	Abusive	**Agressive**	**Hate**	**Identity Hate**	**Insult**	**Neither**	**Obscene**	**Offensive**	**Profane**	**Severe Toxic**	**Threat**	**Toxic**	**Best model**
G^{SUD}	0.77	0.52	0.61	0.11	0.35	0.93	0.11	0.69	0.24	0.30	0.40	0.11	**SVM**
Davidson	–	–	0.31	–	–	0.86	–	0.95	–	–	–	–	GB
Founta	0.89	–	0.37	–	–	0.95	–	–	–	–	–	–	MLR
Fox	–	–	0.54	–	–	0.86	–	–	–	–	–	–	MLR
Gab	–	–	0.89	–	–	0.91	–	–	–	–	–	–	GB
Grimminger	–	–	0.29	–	–	0.96	–	–	–	–	–	–	GB
HASOC2019	–	–	0.24	–	–	0.79	–	0.16	0.40	–	–	–	MLR
HASOC2020	–	–	0.08	–	–	0.89	–	0.36	0.82	–	–	–	SVM
Hateval	–	–	0.63	–	–	0.78	–	–	–	–	–	–	RF
Hateval	–	–	0.66	–	–	0.76	–	–	–	–	–	–	MNB
Hateval	–	–	0.64	–	–	0.77	–	–	–	–	–	–	MLR
Hateval	–	–	0.62	–	–	0.79	–	–	–	–	–	–	GB
Jigsaw	–	–	–	0.32	0.50	0.97	0.26	–	–	0.28	0.53	0.19	SVM
Olid	–	–	–	–	–	0.82	–	0.62	–	–	–	–	SVM
Olid	–	–	–	–	–	0.83	–	0.61	–	–	–	–	MLR
Reddit	–	–	0.76	–	–	0.92	–	–	–	–	–	–	GB
Stormfront	–	–	0.42	–	–	0.95	–	–	–	–	–	–	MLR
Trac	–	0.78	–	–	–	0.61	–	–	–	–	–	–	MNB

5 Experiments

In the following sections, we present the results of our empirical evaluation for the three model families we examined. Throughout all our experiments, we split the datasets, allocating 80% for training, 10% for validation, and 10% for testing purposes. For the sake of reproducibility, we provide the code, and the data used in the experiments along with the respective instructions in an online repository (Niaouri et al. 2024). The repository includes details on the hyperparameters that differ across the models employed in this study, including learning rate, batch size, number of epochs, and the number of layers, among others. For reproducibility purposes, we can provide the saved models, which are substantial in size.

5.1 Shallow Learning Models

We implement our framework employing Natural Language Toolkit (nltk) functions for preprocessing textual data, including tokenization, stop-word removal, lemmatization, and stemming. We then transform processed text data into numerical features using the TextVectorization layer of TensorFlow. In Table 3, we report the performance of the best performing model for each dataset on individual SUD classes.

We observe that MLR consistently emerges as a strong performer across multiple datasets, showcasing its robustness in handling diverse SUD classes. Similarly, GB demonstrates competitive performance, often ranking as the best model on several datasets, while SVM exhibits varying success, with notable achievements in the G^{SUD}, HASOC2020, Jigsaw, and Olid. Consequently, a definitive consensus concerning the best-performing model is lacking in this framework.

The performance variations highlight the algorithm's sensitivity to the characteristics of specific datasets. It is important to note that in a large-scale context, namely in the G^{SUD} dataset, the generalization performance of the models falls short of expectations. The shallow model's ability to discriminate the classes worsens compared to the performance observed on individual datasets.

Table 4: Optimal performing MLM per class and dataset.

	Macro F1 Score												
	Abusive	**Agressive**	**Hate**	**Identity Hate**	**Insult**	**Neither**	**Obscene**	**Offensive**	**Profane**	**Severe Toxic**	**Threat**	**Toxic**	**Best model**
G^{SUD}	0.79	0.64	0.66	0.36	0.50	0.94	0.25	0.75	0.31	0.40	0.43	0.18	**BERT**
G^{SUD}	0.80	0.64	0.66	0.38	0.51	0.94	0.34	0.75	0.33	0.42	0.46	0.20	**ELECTRA**
G^{SUD}	0.80	0.67	0.68	0.42	0.50	0.94	0.25	0.75	0.37	0.42	0.46	0.17	**RoBERTa**
Davidson	–	–	0.46	–	–	0.90	–	0.94	–	–	–	–	ELECTRA
Founta	0.88	–	0.41	–	–	0.95	–	–	–	–	–	–	BERT
Founta	0.88	–	0.42	–	–	0.95	–	–	–	–	–	–	ALBERT
Founta	0.89	–	0.41	–	–	0.96	–	–	–	–	–	–	RoBERTa
Fox	–	–	0.60	–	–	0.79	–	–	–	–	–	–	RoBERTa
Gab	–	–	0.88	–	–	0.91	–	–	–	–	–	–	ALBERT
Gab	–	–	0.89	–	–	0.91	–	–	–	–	–	–	RoBERTa
Grimminger	–	–	0.58	–	–	0.95	–	–	–	–	–	–	ELECTRA
HASOC2019	–	–	0.29	–	–	0.80	–	0.36	0.57	–	–	–	ELECTRA
HASOC2020	–	–	0.22	–	–	0.91	–	0.30	0.83	–	–	–	ELECTRA
Hateval	–	–	0.75	–	–	0.79	–	–	–	–	–	–	ELECTRA
Hateval	–	–	0.75	–	–	0.80	–	–	–	–	–	–	RoBERTa
Jigsaw	–	–	–	0.46	0.57	0.98	0.38	–	–	0.40	0.56	0.30	ELECTRA
Olid	–	–	–	–	–	0.85	–	0.67	–	–	–	–	BERT
Olid	–	–	–	–	–	0.84	–	0.68	–	–	–	–	ELECTRA
Reddit	–	–	0.76	–	–	0.92	–	–	–	–	–	–	ALBERT
Reddit	–	–	0.76	–	–	0.92	–	–	–	–	–		RoBERTa
Stormfront	–	–	0.60	–	–	0.96	–	–	–	–	–	–	RoBERTa
Trac	–	0.81	–	–	–	0.71	–	–	–	–	–	–	BERT

5.2 Masked Language Models

We conduct an experimental evaluation of MLM models using $\text{BERT}_{\text{BASE}}$ (Devlin et al. 2019; Yuan and Rizoiu 2022), pre-trained on text tokenized with the WordPiece algorithm (Wu et al. 2016), and its variants: $\text{ALBERT}_{\text{BASE}}$ (Lan et al. 2019), $\text{RoBERTa}_{\text{BASE}}$ (Liu et al. 2019) and $\text{ELECTRA}_{\text{BASE}}$ (Clark et al. 2020). To perform SUD classification, we connect BERT's 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 the proposed model on the G^{SUD} corpus, adopting a stratified sampling technique to keep the same class distribution throughout the three splits.

Table 4 reports the optimal performing model for each dataset. ELECTRA is shown to be the best performer in most of the corpora as it exhibits the highest Macro F1 score in eight datasets, including G^{SUD}. When comparing the performance of the BERT variants with that of the original BERT model, the results suggest a slightly higher ability of ELECTRA and RoBERTa to discriminate the SUD classes.

5.3 Causal Language Models

Following the methodological steps outlined in the previous section, we conducted fine-tuning on the pre-trained models Llama-2-7b (Touvron et al. 2023), Mistral-7B-v0.1 (Jiang et al. 2023), and mpt-7b (MosaicML NLP team 2023) using our datasets. The fine-tuning procedure employed the Parameter-Efficient Fine-Tuning (PEFT) method, where specific hyperparameters such as learning rates, batch sizes, and adapter weights were configured. Notably, we utilized the SFTTrainer class from the TRL library, designed for training LLMs. Additionally, we created custom prompts to input the categorization of text into the respective classes. For details on the hyperparameters and prompts used see our online repository (Niaouri et al. 2024).

In Table 5, we report the best-performing model for each dataset. The results indicate a notable advantage of the Mistral at a single dataset scale. The second-best performing model, Llama 2, showcases similar results but holds a significant advantage with an F1 score of 41% on the G^{SUD} dataset compared to Mistral's 26% showing that Llama 2 performs better on a larger scale.

Here, the results exhibit similar patterns to the ones observed for the previous models, where we obtain shaky classification results in the hate and offensive classes (majority classes) and low performances in the underrepresented SUD types (i.e., *severe toxic*, *threat*, and *toxic*).

Table 5: Optimal performing CLM per class and dataset.

	Macro F1 Score												
	Abusive	**Agressive**	**Hate**	**Identity Hate**	**Insult**	**Neither**	**Obscene**	**Offensive**	**Profane**	**Severe Toxic**	**Threat**	**Toxic**	**Best model**
G^{SUD}	0.76	0.63	0.30	0	0.48	0.84	0.13	0.32	0.13	0.32	0.36	0.23	**Llama–2–7**
Davidson	–	–	0.45	–	–	0.87	–	0.94	–	–	–	–	Mistral–7B–v0.1
Founta	0.89	–	0.42	–	–	0.91	–	–	–	–	–	–	Mistral–7B–v0.1
Fox	–	–	0.67	–	–	0.82	–	–	–	–	–	–	Mistral–7B–v0.1
Gab	–	–	0.88	–	–	0.89	–	–	–	–	–	–	Llama–2–7b
Gab	–	–	0.89	–	–	0.90	–	–	–	–	–	–	Mistral–7B–v0.1
Grimminger	–	–	0.37	–	–	0.76	–	–	–	–	–	–	Mistral–7B–v0.1
HASOC2019	–	–	0.16	–	–	0.80	–	0.19	0.54	–	–	–	Llama–2–7b
HASOC2020	–	–	0.08	–	–	0.82	–	0.11	0.74	–	–	–	Llama–2–7b
Hateval	–	–	0.76	–	–	0.78	–	–	–	–	–	–	Mistral–7B–v0.1
Jigsaw	–	–	–	0.41	0.53	0.97	0.28	–	–	0.18	0.38	0.10	Llama–2–7b
Olid	–	–	–	–	–	0.85	–	0.64	–	–	–	–	Llama–2–7b
Olid	–	–	–	–	–	0.83	–	0.66	–	–	–	–	mpt–7b
Reddit	–	–	0.77	–	–	0.92	–	–	–	–	–	–	Llama–2–7b
Reddit	–	–	0.78	–	–	0.93	–	–	–	–	–	–	Mistral–7B–v0.1
Stormfront	–	–	0.58	–	–	0.95	–	–	–	–	–	–	Llama–2–7b
Stormfront	–	–	0.61	–	–	0.93	–	–	–	–	–	–	Mistral–7B–v0.1
Trac	–	0.84	–	–	–	0.71	–	–	–	–	–	–	Mistral–7B–v0.1

6 Model comparison and further analyses

In this section, we present the findings from our supplementary evaluations conducted on the optimal model within each model family for the G^{SUD} dataset. We conducted these new experiments in a more controlled environment that allowed us to empirically test our hypothesis on the causes behind the poor generalization performance we observed. We selected ELECTRA for the MLM family due to its superior performance not only on G^{SUD} but also across other datasets, alongside BERT and RoBERTa. For the CLMs, our choice was Llama 2, as it demonstrated a notably higher performance compared to Mistral on the G^{SUD} dataset. Regarding the SLMs, despite the higher performance of SVM on G^{SUD}, MLR was preferred due to enhanced scalability.

Table 6 contains the results for each of the different experimental setups, where we report the Macro F1 score of the SUD classification. Considering that G^{SUD} contains highly unbalanced classes, we repeated classification tasks after training our model on a balanced dataset. Given the dominance of *neither* class, we examined a setting with undersampled non-SUD text (neither class). Hence, we 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 in two out of three model families, with the most noteworthy improvement attested in the SLMs and a significant drop of performance for the CLMs.

Furthermore, we tested the binary classification scenario where models had to differentiate between SUD and non-SUD content, providing a balanced binary setup. To that extent, we performed random oversampling of minority classes as suggested by several works (Yuan and Rizoiu 2022; Swamy, Jamatia and Gambäck 2019; MacAvaney et al. 2019). In this scenario, we achieved a relatively high Macro F1 score (86%, 89%, and 88% for SLMs, MLMs and CLMs respectively) and a tiny improvement when classes were balanced (88%, 90%, and 90%). These outcomes underscore the model's capability to effectively distinguish the *neither* class from generic SUD in the broader contextual framework we built.

A substantial improvement is evident when exclusively assessing the model's performance in a dataset containing only the following classes: *hate*, *offensive*, *toxic*, and *neither*, whose instances are about 90% of the total G^{SUD}. Discriminating over such classes is challenging as they appear in multiple datasets with different annotation schemas.

By removing the *neither* class, and focusing solely on the three specified categories – hate, offensive, and toxic – we aimed to sharpen the analysis of the models' discriminatory power specifically within the SUD classes. This choice was made to

assess whether the models could more effectively differentiate between the nuanced forms of harmful content. The results demonstrated a noteworthy enhancement, reaching a Macro F1 score of 81%, 85%, and 59%, respectively. This rise in performance for the SLMs and MLMs implies that the models can generalize better when the neutral category is absent. Such a scenario indicates a sensitive decrease in false dismissals on the positive SUD classes due to the absence of the neutral class. Conversely, this pattern is not observed within the CLM family, suggesting that the efficacy of Llama 2 is contingent upon the prevalence of the neither class in substantial proportions, as also shown in cases where neither was undersampled.

7 Multi-source learning

In this part, we present the result of our test around the models' capability to learn knowledge from different sources whose labels belong to different annotation schemas. We recall that our main research questions are: Which is the SOTA model generalization capability in a global context, where the models are trained on a general dataset and tested on individual datasets that share some of its classes? What are the main challenges hampering the SUD modeling effectiveness, and how do the different model families perform in a multi-class vs a binary setup? We present the results of our evaluation hereafter.

Table 6: Comparison between all experiments.

	F1 Score	**F1 Score**	**F1 Score**
	SLMs - MLR	**MLMs ELECTRA**	**CLMs Llama-2-7**
Training set	Macro	Macro	Macro
G^{SUD}	0.41	**0.54**	0.41
G^{SUD} with Neither Undersampled	**0.76**	0.60	0.23
G^{SUD} (Binary classification)	0.86	**0.89**	0.88
G^{SUD} balanced (Binary classification)	0.88	**0.90**	**0.90**
G^{SUD} (*hate*, *offensive*, *toxic*, *neither*)	0.63	**0.69**	0.63
G^{SUD} (*hate*, *offensive*, *toxic*)	0.81	**0.85**	0.59

Table 7: Multi-class SUD classification results (F1 score) with the model trained on GSUD vs on each individual dataset.

	Macro F1 Score (%)					
	Multi-class SUD Classification					
	SLMs - MLR		MLMs - ELECTRA		CLMs - Llama-2-7	
Dataset	**Classified in G^{SUD}**	**Individual**	**Classified in G^{SUD}**	**Individual**	**Classified in G^{SUD}**	**Individual**
G^{SUD}	0.41	–	0.53	–	0.41	–
Davidson	0.07	0.70	**0.79**	0.77	0.50	0.73
Founta	0.43	0.74	**0.79**	0.73	0.55	0.74
Fox	0.35	0.70	**0.59**	0.56	0.55	0.65
Gab	0.08	0.89	**0.92**	0.89	0.72	0.89
Grimminger	0.44	0.50	0.72	0.76	**0.57**	0.52
HASOC2019	0.24	0.40	0.45	0.51	**0.44**	0.42
HASOC2020	0.28	0.53	0.54	0.57	0.37	0.44
Hateval	0.51	0.71	0.75	0.78	0.60	0.75
Jigsaw	0.02	0.41	**0.57**	0.52	0.29	0.41
Olid	0.23	0.72	0.74	0.76	0.44	0.75
Reddit	0.09	0.83	**0.85**	0.82	0.71	0.85
Stormfront	0.47	0.68	**0.87**	0.75	0.60	0.77
Trac	0.26	0.69	**0.86**	0.75	0.57	0.76

7.1 Multi-source learning in multi-class SUD classification

In Table 7, we depict the classification results obtained for each dataset by models trained on the (large-scale) G^{SUD} corpus compared to the models trained in each dataset.

We note that, almost exclusively in the MLM family and in ~50% of the cases (highlighted in bold), the model trained on G^{SUD} is slightly better than the specialized counterpart. This outcome allows us to conclude that leveraging more knowledge from multiple domains has several advantages despite different dataset incongruences observed in the previous experiments.

Table 8: Binary SUD classification with the models trained in GSUD.

	Macro F1 Score (%)		
	Binary SUD Classification – Classified in G^{SUD}		
Dataset	**SLMs-LR**	**MLMs - ELECTRA**	**CLMs - Llama-2-7**
G^{SUD}	0.86	**0.89**	0.88
Davidson	0.06	**0.96**	0.79
Founta	0.15	**0.95**	0.86
Fox	0.51	**0.79**	0.49
Gab	0.15	**0.89**	0.78
Grimminger	0.60	**0.85**	0.58
HASOC2019	0.58	**0.82**	0.53
HASOC2020	0.82	**0.95**	0.78
Hateval	0.66	**0.79**	0.59
Jigsaw	0.06	**0.93**	0.74
Olid	0.22	**0.87**	0.62
Reddit	0.19	**0.81**	0.74
Stormfront	0.67	**0.86**	0.64
Trac	0.35	**0.88**	0.38

7.2 Multi-source learning in binary SUD classification

For each of the experiments reported in this section, we have also tested the capability of the models to discriminate SUD and non-SUD text in G^{SUD} (depicted in Table 8).

Here, we obtain a relatively high Macro F1 score (~90%) under the G^{SUD} condition, noticing that the models discriminate well the *neither* class from the generic SUD in the global context we built. Such results confirm the current trend observed in the ML literature so far (e.g., Swamy et al. 2019; Antypas and Camacho-Collados 2023). Concerning the generalization capabilities of the models, the MLM family seems to be the best-performing one, followed by the CLMs. However, the generalization capability of the SLM models is low, as seen from the performances attested for each of the individual datasets.

8 Model explainability

Here, we focus on the most effective model family, namely the MLMs. We aim to explain several aspects regarding the performance of ELECTRA, examining the relationship between the model's capability to distinguish SUD classes and the impact of balanced datasets on classification performance. To clarify the discriminative capacity of the adopted model across SUD classes, we employ a visualization technique on the generated text representation, specifically on ELECTRA's pooled output layer. We reduce the output dimensionality (2 dimensions) using t-distributed Stochastic Neighbor Embedding (t-SNE). The resulting plot, as shown in Figure 2, illustrates the outcomes of the test set under two distinct training scenarios: (a) the model trained on the complete G^{SUD} corpus, and (b) the model trained on the G^{SUD} with the *neither* class being undersampled.

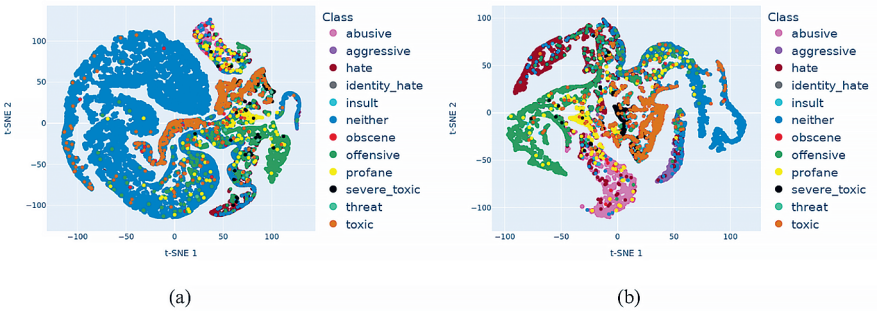

Figure 2: Two dimensional t-SNE visualization of sample embedding produced by ELECTRA's pooled output layer: (a) G^{SUD} (b) G^{SUD} with neither undersampled.

Observations from Figure 2(a) reveal that certain classes, such as *abusive* (top-right) and toxic (center-right), form distinct clusters, indicating a clear separation in the data. This behavior reflects the exclusive occurrence of these classes in the individual datasets, as evidenced in Table 4. Conversely, classes like *profane*, *obscene*, *threat*, and *severe toxic*, are distributed throughout the plot and are not easily distinguishable. Overall, we note that low performances are observed not only in classes with minimal training samples but also in those sharing samples from multiple corpora, indicating the presence of heterogeneous intraclass samples. The *hate* class represents a notable example, encompassing samples from ten datasets (out of thirteen).

In Figure 2(b), where *neither* class is undersampled, we observe a notable enhancement in clustering quality. This is evidenced by the improved performance

in F1 score, as illustrated in Table 6, highlighting the model's more accurate classification under balanced conditions. Notably, the clusters representing the *abusive* and *aggressive* classes are easily distinguishable, further confirming that the model can more accurately classify classes originating from a single dataset. Next, we observe that the classes for *hate*, *offensive*, and *toxic* content also form distinct clusters, although some outliers are still present. Finally, the categories of *profane*, *severe toxic*, *threat*, *insult*, and *obscene* content remain more scattered.

In Figure 3, we visualize the embeddings produced by models trained on (a) G^{SUD} with *hate*, *offensive*, *toxic*, and *neither* classes and (b) G^{SUD} with *hate*, *offensive*, and *toxic* classes excluding the *neither* class.

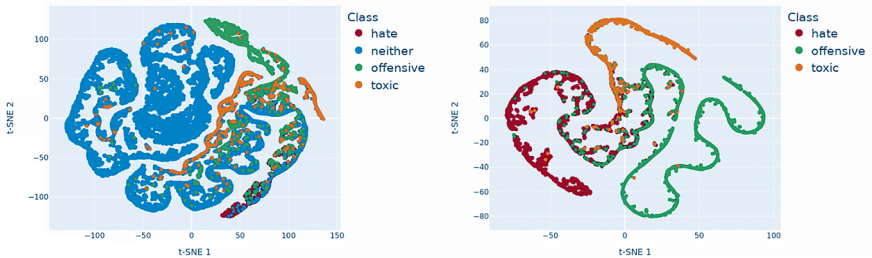

Figure 3: Two components t-SNE visualization of samples embedding produced by ELECTRA's pooled output layer: (a) G^{SUD} (hate, offensive, toxic, neither), (b) G^{SUD} (hate, offensive, toxic).

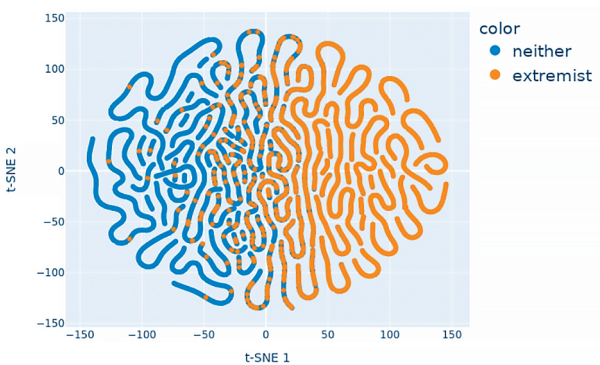

Figure 4: Two components t-SNE visualization of samples embedding produced by ELECTRA's pooled output layer: Binary Classification.

Figure 3(a) validates the hypothesis that a less overlapping annotation schema yields more promising results and better-defined clusters. Figure 3(b) demonstrates the model's ability to differentiate SUD classes even when *neither* class is absent, confirming its crucial (negative) role in the multi-class model fine-tuning.

Finally, in Figure 4, we present the binary classification case, emphasizing the high discriminative power of ELECTRA, which in this problem setting (simpler than multi-class) can separate the search space with high accuracy.

9 Discussion

Our study delves into assessing the effectiveness of three distinct categories of language models, namely Shallow Learning Models (SLMs), Masked Language Models (MLMs), and Causal Language Models (CLMs), in classifying SUD. Within the SLMs, MLR consistently demonstrated the best performance. Among the MLMs, ELECTRA exhibited superior efficacy, while within the CLM family, Mistral showed its superiority in individual datasets but fell short in the G^{SUD} dataset. Balanced dataset configurations and binary classification scenarios enhanced model performance, underscoring the significance of clearly defined class boundaries and balanced training data. This result is expected, given that various studies have highlighted the advantages of balanced datasets in hate speech classification (Qureshi and Sabih 2021) and the improved performance of models in binary rather than multi-class settings (Bouazizi et al. 2016). Our findings further suggest that inadequate training samples and intraclass variability – where a class encompasses a diverse range of samples from multiple sources – can negatively impact model performance. Current SOTA models for SUD classification require consistent dataset annotations and homogeneous samples to optimize classification performance.

Another significant finding from our study, which focused on large-scale multi-source learning, is that MLMs displayed superior generalization capabilities compared to the other two model families followed by the CLMs that demonstrated comparable performances yet exhibited difficulties in the experimental conditions where the *neither* class was either undersampled or absent.

The superiority of MLMs in SUD classification stems from their bidirectional context awareness, as at the training stage, they consider both preceding and following tokens. This characteristic has consistently led to better performance in comparison to Shallow Learning approaches, as demonstrated by models like BERT, RoBERTa, and ALBERT (Swamy, Jamatia and Gambäck 2019; Markov and Daelemans 2021; Fortuna, Soler-Company and Wanner 2021). Notably, ELECTRA stands out in G^{SUD} and individual datasets due to its distinctive architecture, which involves

training a generator whose task is to replace sentence tokens and a discriminator that learns to identify the replaced token. Several studies have highlighted ELECTRA's efficacy in various classification and sentiment analysis tasks, and its performances often remain closely aligned with other BERT variants (Guven 2021; Pedersen et al. 2022; Kowsher 2023). While ELECTRA offers certain advantages, the overall effectiveness of MLMs is robust across different architectures. Another advantage of the MLM model family is the significantly faster learning task concerning CLMs. In our case, on the G^{SUD} dataset, CLMs required up to a week to complete tasks, whereas the slower MLM learning process took less than 24 hours. We observe a similar trend in smaller datasets, where MLMs completed tasks in a few hours, while CLMs took several days. Among all model families, SLMs generally exhibited the fastest running time, except for the SVM model.

10 Conclusion and future work

In this work, we present an empirical evaluation of automatic SUD detection using a variety of models constructing a comprehensive framework of SOTA solutions for SUD classification. To test generalization capability, we considered a large and heterogeneous context in which we obtained varying results, not always in line with the expected performance of the model trained at the local level, i.e., on every individual 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.).

For 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 additionally highlight the significant potential of our findings for researchers in linguistics, discourse analysis, and 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. 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 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.

Acknowledgments

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.

This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD010615085R1 made by GENCI.

References

Antypas, Dimosthenis & Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. *The 7th Workshop on Online Abuse and Harms (WOAH).* Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.woah-1.25.

Alkomah, Fatimah & Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information* 13 (6). 273. https://doi.org/10.3390/info13060273.

Aroyehun, Segun Taofeek, & Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. *TRAC@COLING 2018*. https://aclanthology.org/W18-4411 (last accessed 14 February 2025).

Ascone, Laura, & Julien Longhi. 2018. The expression of threat in jihadist propaganda. *Fragmentum* 50. 85. https://doi.org/10.5902/2179219428823.

Badjatiya, Pinkesh, Manish Gupta & Vasudeva Varma. 2019. Stereotypical bias removal for hate Speech detection task using knowledge-based generalizations. *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313504.

Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta & Vasudeva Varma. 2017. Deep learning for hate speech detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion.* https://doi.org/10.1145/3041021.3054223.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso & Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual detection of hate pseech against immigrants and women in Twitter. *International Workshop on Semantic Evaluation*. https://doi.org/10.18653/v1/s19-2007.

Bouazizi, Mondher & Tomoaki Ohtsuki. 2016. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. *2016 IEEE International Conference on Communications (ICC)*. https://doi.org/10.1109/icc.2016.7511392.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45 (1). 5–32. https://doi.org/10.1023/a:1010933404324.

Clark, Kevin, Minh-Thang Luong, Quoc Le V & Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2003.10555.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.747.

Davidson, Thomas, Dana Warmsley, Michael Macy & Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media* 11 (1). 512–515. https://doi.org/10.1609/icwsm.v11i1.14955.

Davidson, Thomas, Debasmita Bhattacharya & Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *Proceedings of the Third Workshop on Abusive Language Online*. https://doi.org/10.18653/v1/w19-3504.

De Giorgio, Andrea, Goran Kuvačić, Dražen Maleš, Ignazio Vecchio, Cristina Tornali, Wadih Ishac, Tiziana Ramaci, Massimiliano Barattucci & Boris Milavić. 2022. Willingness to receive COVID-19 booster vaccine: Associations between green-pass, social media information, anti-vax beliefs, and emotional bBalance. *Vaccines* 10 (3). 481. https://doi.org/10.3390/vaccines10030481.

De Gibert, Ona, Naiara Perez, Aitor García-Pablos & Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *ArXiv, Abs/1809.04444*. https://doi.org/10.18653/v1/w18-5102.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/n19-1423.

Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. *ALW@ACL*. https://doi.org/10.18653/v1/w17-3007.

Fortuna, Paula, Juan Soler-Company & Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* 58 (3). 102524. https://doi.org/10.1016/j.ipm.2021.102524.

Founta, Antigoni, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos & Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media* 12 (1). https://doi.org/10.1609/icwsm.v12i1.14991.

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29 (5). https://doi.org/10.1214/aos/1013203451.

Gandhi, Ankita, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria & Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems* 41 (8). https://doi.org/10.1111/exsy.13562.

Gao, Lei & Ruihong Huang. 2017. Detecting online hate speech using context aware models. *ArXiv, Abs/1710.07395*. https://doi.org/10.26615/978-954-452-049-6_036.

Grimminger, Lara & Roman Klinger. 2021. Hate towards the political opponent: A Twitter Corpus study of the 2020 US elections on the basis of offensive speech and stance Detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.01664.

Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Conti & N. Asokan. 2018. All you need is "love." *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. https://doi.org/10.1145/3270101.3270103.

Guven, Zekeriya Anil. 2021. The effect of BERT, ELECTRA and ALBERT language models on sentiment analysis for Turkish product reviews. *2021 6th International Conference on Computer Science and Engineering (UBMK)*. https://doi.org/10.1109/ubmk52708.2021.9559007.

Hearst, M.A., S.T. Dumais, E. Osuna, J. Platt & B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and Their Applications* 13 (4). 18–28. https://doi.org/10.1109/5254.708428.

Janiesch, Christian, Patrick Zschech & Kai Heinrich. 2021. Machine learning and deep learning. *Electronic Markets* 31 (3). 685–695. https://doi.org/10.1007/s12525-021-00475-2.

Jasser, Greta, Jordan McSwiney, Ed Pertwee & Savvas Zannettou. 2021. 'Welcome to #GabFam': Far-right virtual community on Gab. *New Media & Society* 25 (7). 1728–1745. https://doi.org/10.1177/1461444 8211024546.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego De Las Casas, Florian Bressand, et al. 2023. Mistral 7B. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06825.

Karan, Mladen & Jan Šnajder. 2018. Cross-Domain Detection of Abusive Language Online. *Workshop on Abusive Language Online*. https://doi.org/10.18653/v1/w18-5117.

Kibriya, Ashraf M., Eibe Frank, Bernhard Pfahringer & Geoffrey Holmes. 2004. Multinomial naive bayes for text categorization revisited. *In Lecture notes in computer science* 488–499. https://doi.org/ 10.1007/978-3-540-30549-1_43.

Kowsher, Md. 2023. Analyzing the impact of transfer learning from pretrained transformers on text classification: A cross-model study. *ResearchGate*. https://doi.org/10.13140/RG.2.2.29289.26729/1.

Kumar, Ritesh, Aishwarya N. Reganti, Akshit Bhatia & Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. *ACL Anthology.* https://aclanthology.org/L18-1226/ (last accessed 14 February 2025).

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma & Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv.org.* http://www.arxiv.org/abs/1909.11942 (last accessed 14 February 2025).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. ROBERTA: A robustly optimized BERT pretraining approach. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.1907.11692.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian & Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE* 14 (8). e0221152. https://doi.org/ 10.1371/journal.pone.0221152.

Mandl, Thomas, Sandip Modha, Anand Kumar M & Bharathi Raja Chakravarthi. 2020. Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. *Forum for Information Retrieval Evaluation.* https://doi.org/10.1145/ 3441501.3441517.

Markov, Ilia & Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. https://doi.org/10.18653/v1/2021.nlp4if-1.3.

Modha, Sandip, Thomas Mandl, Prasenjit Majumder, Daksh Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. *Conference-proceeding*. https://ceur-ws.org/Vol-2517/T3-1.pdf (last accessed 14 February 2025).

MosaicML NLP Team. 2023. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs. *Databricks*. https://www.databricks.com/blog/mpt-7b.

Neyshabur, Behnam, Hanie Sedghi & Chiyuan Zhang. 2020. What is being transferred in transfer learning? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2008.11687.

Niaouri Dimitra, Bruno Machado Carneiro, Michele Linardi, and Julien Longhi (2024). Machine _ learning_heading_to_SUD. https://github.com/diniaouri/Machine_Learning_heading_to_SUD (last accessed 14 February 2025).

Pamungkas, Endang Wahyu & Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. A*nnual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/p19-2051.

Pahor De Maiti, Kristina, Darja Fišer, Nikola Ljubešić, Tomaž Erjavec. 2020. Grammatical footprint of socially unacceptable Facebook comments. Journal-article. *FRENK Corpus*. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_PahordeMaiti-et-al_Grammatical-Footprint-of-Socially-Unacceptable-Facebook-Comments.pdf (last accessed 14 February 2025).

Pedersen, Jannik S., Martin S. Laursen, Cristina Soguero-Ruiz, Thiusius R. Savarimuthu, Rasmus Sogaard Hansen & Pernille J. Vinholt. 2022. Domain over size: Clinical ELECTRA surpasses general BERT for bleeding site classification in the free text of electronic health records. *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. https://doi.org/10.1109/bhi56158.2022.9926955.

Piot, Paloma, Patricia Martín-Rodilla & Javier Parapar. 2024. MetaHate: A dataset for unifying efforts on hate speech detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.06526.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco & Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation* 55 (2). 477–523. https://doi.org/10.1007/s10579-020-09502-8.

Postigo-Fuentes, Ana Yara, Rolf Kailuweit, Alexander Ziem, Stefan Hartmann. 2024. Defining extremist narratives: A review of the current state of the art. In *HORIZON – CL2-2022-DEMOCRACY -01-05* [Report]. Momentum Consulting. https://arenasproject.eu/download/1545/?tmstv=1721660114 (last accessed 14 February 2025).

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding & William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *Conference on Empirical Methods in Natural Language Processing.* https://doi.org/10.18653/v1/d19-1482.

Qureshi, Khubaib Ahmed & Muhammad Sabih. 2021. Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access* 9. 109465–109477. https://doi.org/10.1109/access.2021.3101977.

Röttger, Paul, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts & Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. *arXiv Preprint arXiv:2012.15606.* https://doi.org/10.18653/v1/2021.acl-long.4.

Roy, Sayar Ghosh, Ujwal Narayan, Tathagata Raha, Zubair Abid & Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2101.03207.

Salminen, Joni, Maximilian Hopf, Shammur A. Chowdhury, Soon-Gyo Jung, Hind Almerekhi & Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10 (1). https://doi.org/10.1186/s13673-019-0205-6.

Sulc, Ajda & Kristina Pahor De Maiti. 2020. No room for hate: What research about hate speech taught us about collaboration? https://www.semanticscholar.org/paper/No-room-for-hate%3A-What-research-about-hate-speech-Sulc-Maiti/b04049a663c7c91dc87a16d4a179073861978798 (last accessed 14 February 2025).

Swamy, Steve Durairaj, Anupam Jamatia & Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).* https://doi.org/10.18653/v1/k19-1088.

Toraman, Cagri, Furkan Şahinuç & Eyup Halit Yılmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2203.01111.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. LLAMA: Open and efficient foundation language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971.

Van Aken, Betty, Julian Risch, Ralf Krestel & Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *ArXiv, Abs/1809.07572*. https://doi.org/10.18653/v1/w18-5105.

Wang, Bin, Yunxia Ding, Shengyan Liu & Xiaobing Zhou. 2019. YNU_Wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. https://www.semanticscholar.org/paper/YNU_Wb-at-HASOC-2019%3A-Ordered-Neurons-LSTM-with-for-Wang-Ding/421b9e3f18202b757f0de42ca4a1d2de7dbe29ba (last accessed 14 February 2025).

Wright, Raymond E. 1995. Logistic regression. In L. G. Grimm & P. R. Yarnold (eds.), Reading and understanding multivariate statistics. 217–244. *American Psychological Association*. https://www.scirp.org/reference/referencespapers?referenceid=3007390 (last accessed 14 February 2025).

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.1609.08144.

Xu, Yayin, Ying Zhou, Przemyslaw Sekula & Lieyun Ding. 2021. Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment* 6. 100045. https://doi.org/10.1016/j.dibe.2021.100045.

Yigezu, M., O. Kolesnikova, G. Sidorov & A. Gelbukh. 2023. Transformer-Based hate speech detection for multi-class and multi-label classification. *IberLEF@SEPLN*. https://www.semanticscholar.org/paper/Transformer-Based-Hate-Speech-Detection-for-and-Yigezu-Kolesnikova/165e336c9c6780fad66a68b7be938593b4221149 (last accessed 14 February 2025).

Yin, Wenjie & Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science* 7. e598. https://doi.org/10.7717/peerj-cs.598.

Yu, Zehui, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza & Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review.* https://doi.org/10.1177/08944393241258771.

Yuan, Lanqin & Marian-Andrei Rizoiu. 2022. Detect hate speech in unseen domains using multi-task learning: A case study of political public figures. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2208.10598.

Yuan, Lanqin, Tianyu Wang, Gabriela Ferraro, Hanna Suominen & Marian-Andrei Rizoiu. 2023. Transfer learning for hate speech detection in social media. *Journal of Computational Social Science* 6 (2). 1081–1101. https://doi.org/10.1007/s42001-023-00224-9.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/n19-1144.