[MASK]ED - Language Modeling for Explainable Classification and Disentangling of Socially Unacceptable Discourse.

Dimitra Niaouri[1*], Mohamed Rayane Ghilene[2*], Michele Linardi[1], Julien Longhi[2]

[1]ETIS UMR-8051, CY Cergy Paris University, [2]AGORA, CY Cergy Paris University

Correspondence: dimitra.niaouri@cyu.fr, mohamed-rayane.ghilene@cyu.fr, michele.linardi@cyu.fr, julien.longhi@cyu.fr

Abstract

Analyzing Socially Unacceptable Discourse (SUD) online is a critical challenge for regulators and platforms amidst growing concerns over harmful content. While Pre-trained Masked Language Models (PMLMs) have proven effective for many NLP tasks, their performance often degrades in multi-label SUD classification due to overlapping linguistic cues across categories. In this work, we propose an artifact-guided pre-training strategy that injects statistically salient linguistic features, referred to as artifacts, into the masked language modeling objective. By leveraging context-sensitive tokens, we guide an importance-weighted masking scheme during pre-training to enhance generalization across discourse types. We further use these artifact signals to inform a lightweight dataset curation procedure that highlights noisy or ambiguous instances. This supports targeted relabeling and filtering, enabling more explainable and consistent annotation with minimal changes to the original data. Our approach provides consistent improvements in 10 datasets extensively used in SUD classification benchmarks.

Disclaimer: This article contains some extracts of unacceptable and upsetting language.

1 Introduction

In an era defined by global crises, rising inequality, and the proliferation of extreme online content, regulators at different levels pressingly need to adopt effective Machine Learning (ML) solutions to detect Socially Unacceptable Discourse (SUD). The ever-changing and evolving nature of social discourse not only presents significant challenges to understanding it but also limits the capabilities of the available discourse analysis tools. Analyzing SUD on the other hand is an even more challenging task that requires context-aware models capable of understanding its subtleties and nuances.

Pre-trained Masked Language Models (PMLMs) have proven effective in different NLP tasks, including accurate classification of inadequate content (Swamy et al., 2019a; Markov and Daelemans, 2021a; Fortuna et al., 2021; Yin and Zubiaga, 2021; Toraman et al., 2022; Antypas and Camacho-Collados, 2023; Yigezu et al., 2023; Carneiro et al., 2023; Niaouri et al., 2025). These models however, face multiple challenges when used for online discourse analysis (Carneiro et al., 2023; Ghilene et al., 2024; Niaouri et al., 2025), where they require to learn from noisy data containing multiple distributions annotated with shallow categories (Niaouri et al., 2024). In this scenario, counting on inaccurate content labelling can in turn lead to severe (or too weak) censorships that may disadvantage content creators or contributing to information gaps (Draper and Neschke, 2023). We also notice that PMLMs under-perform when required to generalize over different label distributions (multi-class) and thus have to specialize over different types of speech that hereafter we refer to as SUD (Vehovar et al., 2020; de Maiti and Fišer, 2021; Carneiro et al., 2023).

Masked Language Models (MLMs) are often trained with a random masking schema over a generic corpus, where a model learns to predict randomly masked (and/or replaced) tokens considering their surrounding context (Devlin et al., 2019).

Intuitively, we can expect that the same linguistic pattern or keywords can appear in different types of discourse, breaking the assumption that a given class has a unique structure and vocabulary. In such a case, the language model ability to recognize and disentangle a given language feature depends on the model's understanding of the context around recurrent textual fragments that statistically represent a given class. To discover such patterns, an MLM must learn the heterogeneous contexts surrounding the pivotal textual feature. This begs the question:

[*]These authors contributed equally.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14870–14883

Could we improve the performance of MLMs by selectively focusing on more representative SUD tokens/features during the pre-training?

Recent research efforts (Ramponi and Tonelli, 2022; Levine et al., 2020; Moon et al., 2020) have highlighted the advantage of leveraging relevant textual features to enhance the language model generalization benefits to downstream task such as SUD classification.

We thus study the possibility of injecting statistical knowledge of SUD features at the MLMs pre-training stage. Furthermore, such a strategy permits us to obtain interpretable cues over the model decision space that we can also leverage to estimate noisy labels and perform data curation over the analyzed corpora. We notice that such a strategy is so-far overlooked in the automatic SUD analysis literature.

Contribution: In this work we propose a new SUD classification framework that selectively focuses on and ranks informative tokens related to SUD categories. In turn, it leverages such a knowledge to train unsupervised and supervised MLMs. The proposed approach uses the contextual significance of tokens, to weight the training loss and to estimate noisy labels. We extensively evaluate our contribution in building a SUD classification benchmark with 13 different datasets.

2 Related Work

SUD classification Recent works have advanced various techniques for detecting hate speech, which represent important components of SUD analysis. These methods range from traditional machine learning classifiers to modern transformer-based language models.

A wide range of annotated datasets has been developed to support this task (Davidson et al., 2017; Gao and Huang, 2017; Founta et al., 2018; Van Aken et al., 2018; Kumar et al., 2018; De Gibert et al., 2018; Qian et al., 2019; Basile et al., 2019; Zampieri et al., 2019; Mandl et al., 2019, 2020; Grimminger and Klinger, 2021; Kocon et al., 2021; Yuan and Rizoiu, 2022; Hartvigsen et al., 2022; Mollas et al., 2022), with hatespeechdata[1] serving as a centralized resource for such datasets. Increasing attention has been also devoted to cross-dataset and cross-domain generalization (Karan and Šnajder, 2018; Swamy et al., 2019b; Pamungkas and Patti, 2019; Salminen et al., 2020; Markov and Daelemans, 2021b; Fortuna et al., 2021; Yin and Zubiaga, 2021; Toraman et al., 2022; Antypas and Camacho-Collados, 2023; Malik et al., 2024), multilingual detection (Pamungkas and Patti, 2019; Toraman et al., 2022; Malik et al., 2024), and multiclass or multi-label classification targeting fine-grained categories and underrepresented populations (Toraman et al., 2022; Antypas and Camacho-Collados, 2023; Yigezu et al., 2023; Gandhi et al., 2024).

In this context, Carneiro et al. (2023) and Niaouri et al. (2025) introduced a novel corpus combining texts from diverse platforms (social media, forums, news comments) and annotation schemes. Building on this resource, Niaouri et al. (2025) conducted a comprehensive benchmark of state-of-the-art methods, exposing inconsistencies in existing datasets, such as divergent contextual interpretations of the same labels, and highlighting annotation biases, showing that models trained on single-domain data can experience performance drops of up to 28% in cross-domain evaluations.

Explainable Hate Speech detection Several explainable hate speech detection efforts have focused on developing frameworks that combine detection accuracy with interpretable reasoning.

Hartvigsen et al. (2022) introduce explainable models that leverage narrative structures to capture subtle, implicit hateful expressions and adversarial language, while also releasing ToxiGen, a large-scale machine-generated dataset for implicit hate speech detection. The HARE framework (Yang et al., 2023b) complements this by using Large Language Models (LLMs) with chain-of-thought prompting to generate detailed rationales that improve detection accuracy and explanation quality. By addressing logical inconsistencies across human-annotated datasets, HARE achieves both improved detection performance (3.8% F1 increase over baselines) and enhanced generalizability.

A complementary direction of work enriches explanations with external knowledge and toxicity attributes. For example, Sridhar and Yang (2022) present a knowledge-informed encoder–decoder framework that integrates multiple sources of external knowledge to generate nuanced explanations of stereotypes in biased language, outperforming prior state-of-the-art approaches. Similarly, Yadav et al. (2024) propose Tox-BART, a generative model that leverages explicit toxicity attributes to interpret implicit hate speech.

[1]https://hatespeechdata.com/

Benchmark corpora have been critical in advancing this line of research. The Social Bias Frames dataset (SBIC) (Sap et al., 2020) emphasizes reasoning about the pragmatic frames in which social biases and stereotypes are projected, while the LatentHatred dataset (ElSherief et al., 2021) provides fine-grained annotations of implicit hate speech based on a six-class taxonomy, supporting evaluation of models on subtle and indirect forms of toxicity. Likewise, the HateXplain dataset (Mathew et al., 2022) incorporates word- and phrase-level human rationales across 20K posts, enabling simultaneous assessment of classification accuracy and explanation faithfulness, and revealing that high-performing models often diverge from human reasoning.

Beyond dataset-driven advances, explainable detection has also leveraged deep learning architectures in conjunction with post-hoc interpretability methods, such as attention and attribution analyses, which provide insights into model behavior (Murad et al., 2024).

In contrast to these approaches, our work focuses on mitigating bias and improving cross-domain generalization (without requiring additional annotations), offering interpretable cues tied to model confidence in large-scale, heterogeneous settings where label noise is prevalent.

3 Proposed Approach

This section outlines our methodology for enhancing SUD classification. In Figure 1, we show the overall architecture of our SUD classification framework composed by three novel blocks : I) Token Scoring, II) Pre-training strategy based on weighted loss according important tokens (a.k.a. artifacts), III) Post classification Token Diagnostic that supports Label Noise detection and Dataset Relabelling.

In the following parts, we provide a linguistically grounded definition of SUD describing the details of each architecture building block.

3.1 Socially Unacceptable Discourse (SUD)

Definition Socially Unacceptable Discourse (SUD) encompasses a spectrum of harmful communicative acts characterized by offensive, inciting, or derogatory language. This includes both explicit and implicit threats, negative stereotyping, obscene expressions, and aggressive or dehumanizing rhetoric (Vehovar et al., 2020; de Maiti and Fišer, 2021). From a linguistic standpoint, SUD often parallels hate speech and extremist narratives, exhibiting features such as objectifying nominalizations, third-person plural pronouns that reinforce in-group/out-group dynamics, present-tense constructions that create immediacy, and imperative verbs that encourage harmful behavior (Okulska and Kołos, 2024).

SUD Classification Given a text item, namely a sequence of T tokens $X = (x_1, x_2, \dots, x_T)$, SUD classification assigns to X one of K predefined categories $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$, each corresponding to a distinct type of harmful or inappropriate discourse labelled in the corpus.

A fundamental challenge in SUD classification lies in its context-dependence: the same lexical items may function differently across discourse types generating noisy labels in widely used datasets that solely dispose of context-insensitive annotations. To address this, we propose a context-aware artifact extraction (Figure 1(a)) that selectively emphasizes semantically informative tokens used during bi-directional transformer model pre-training (Figure 1(b)). Extracting and scoring language artifacts will first permit us to weigh their importance, raising the focus on statistically relevant contexts that, in turn, we leverage to define an explainable methodology and estimate label noise in SUD annotated corpora.

We observe that not every token in a text sequence contributes equally to semantic richness or contextual understanding. In this sense, model training by random token masking (Meng et al., 2024), while widely adopted, can result in suboptimal representation learning by overemphasizing frequent but uninformative tokens. In the following part, we describe our artifact scoring and extraction method.

3.2 Token Scoring

PMI Importance Score To estimate token salience across SUD categories, we adopt a variant of Pointwise Mutual Information (PMI) (Fano, 1963), a well-established measure of word-class association. Following Gururangan et al. (2018) and Ramponi and Tonelli (2022), we compute:

$$PMI(x_t, c) = \log_2 \left(\frac{P(x_t|c)}{P(x_t) \cdot P(c)} \right) \quad (1)$$

where $P(x_t|c)$ is the conditional probability of token x_t given its context class c, $P(x_t)$ its marginal

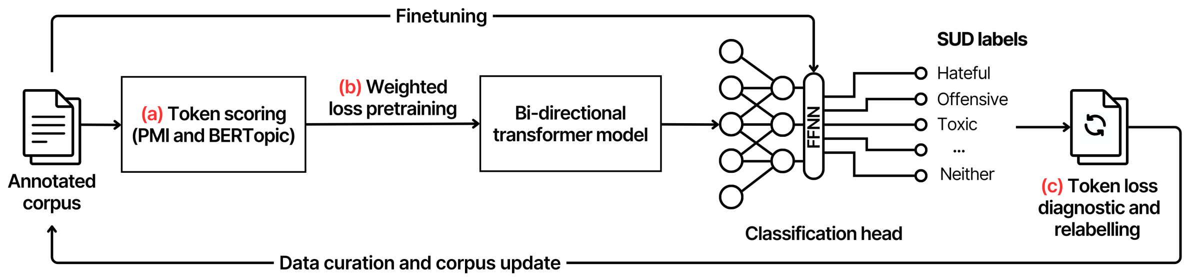

Figure 1: SUD Classification Framework. *(a)* Token Scoring and ranking for each instance in the corpus. *(b)* SUD Artifact-weighted pretraining of a Bi-directional transformer at the core of the SUD classifier (Feed Forward Neural Network - FFNN + Softmax Layer). *(c)* Post-classification Token diagnostic and Noise-Driven score computation for data curation.

probability in the overall corpus, and $P(c)$ the prior probability of class c. To improve comparability and mitigate sensitivity to low-frequency tokens, we normalize PMI using the normalized PMI score (NPMI), and further rescale it to the range $[0, 1]$ using min-max normalization. When a token appears across multiple classes, we compute its overall importance score as the average of its scaled NPMI scores across all associated classes.

BERTopic (BT) importance Score Our second extraction strategy employs BERTopic (Grootendorst, 2022), a transformer-based topic modeling framework that clusters semantically similar texts using Sentence-BERT embeddings [2] (Reimers, 2019). We identify salient tokens within each topic using class Term Frequency - Inverse Document Frequency (cTF-IDF) (Joachims et al., 1997). This metric reflects a token's frequency within a topic relative to its frequency across the corpus:

$$cTF\text{-}IDF(t, T_i) = P(t \mid T_i) \cdot \log\left(\frac{N}{|D_t|}\right) \quad (2)$$

where $P(t \mid T_i)$ is the normalized frequency of token t in its topic $T_i \in \mathbb{N}$, N is the total number of documents, and $|D_t|$ is the number of documents containing t. In our work, we consider that a token is assigned to one or multiple topics in an unsupervised manner using a clustering algorithm (Grootendorst, 2022). Hence, T_i represents a cluster index. A global importance score is then obtained by averaging each token's normalized relevance across all topics in which it appears.

[2]We use the `paraphrase-MiniLM-L3-v2` model to generate sentence embeddings.

3.3 Artifact-Guided Pre-training

Extracting and scoring tokens permits us to incorporate an importance score into the MLM pretraining objective. This integration biases the loss function, assigning a weight to the selected tokens. In this section, we present the masking strategy we adopt and the details of the aforementioned loss function.

Token Masking Strategies To investigate the effect of artifact-guided masking, we consider four masking strategies : **(1) Random Masking**, in which tokens are masked uniformly at random, and the standard MLM loss is applied (Devlin et al., 2019). **(2) Top-k Masking**, where only the k percentage tokens, with the highest importance scores are masked during training. **(3) Random Masking with Weighted Loss**, where tokens are randomly masked, while the loss is scaled by token-level importance weights. **(4) Top-k Masking with Weighted Loss**, extends the Top-k Masking approach by applying an importance-weighted loss to the masked tokens.

Weighted MLM Objective As depicted in Figure 2, given a corpus $\mathcal{D} = \{X_1, \ldots, X_N\}$, where each text item is composed by a set of tokens, namely $X_i = (x_1, \ldots, x_T)$, a subset $\mathbf{M} \subseteq \{1, \ldots, T\}$ is selected for masking. Tokens at these positions are replaced with the [MASK] string placeholder. The model is trained to predict each masked token x_t from the corrupted sequence $\tilde{X}_i$, minimizing the standard negative log-likelihood:

$$\ell_t = -\log\left(p_\theta(x_t \mid \tilde{X})\right) \quad (3)$$

To emphasize semantically salient tokens, each ℓ_t is scaled by an importance score w_t, derived from our artifact extraction methods presented in Section

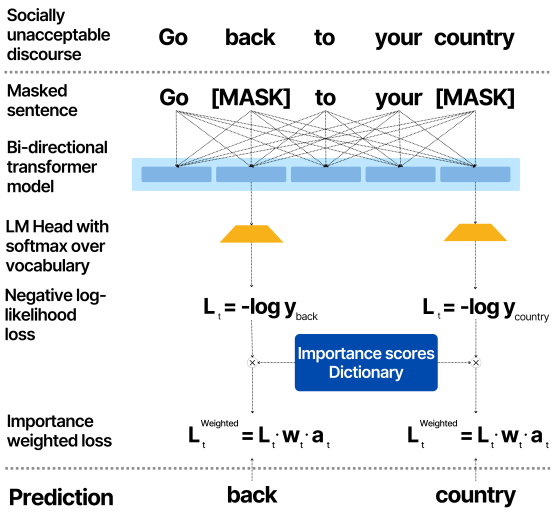

Figure 2: Masked language model pre-training. Tokens are passed through a bidirectional encoder model (Devlin et al., 2019). The token prediction loss is then weighted using the pre-computed importance scores.

3.2 and a binary mask indicator ($a_t = \mathbf{1}[t \in \mathbf{M}]$):

$$\ell_t^{\text{Weighted}} = \ell_t \cdot w_t \cdot a_t \tag{4}$$

The final objective averages the weighted loss across all masked positions:

$$\mathcal{L}_{\text{MLM}}^{\text{Weighted}} = \frac{\sum_{t=1}^{T} \ell_t^{\text{Weighted}}}{\sum_{t=1}^{T} a_t} \tag{5}$$

This formulation biases training toward artifact-relevant tokens while maintaining stable optimization across variable-length sequences.

Framework Architecture Figure 2 illustrates with an example the artifact-guided masked language model pre-training. Masked tokens are processed through a stack of bidirectional transformer layers, yielding contextualized representations h_t^L. These hidden states are projected into the vocabulary space using a learned output matrix E, producing logits $u_t = h_t^L E^\top$, which are then transformed into output probabilities via a softmax function:

$$y_t = \text{softmax}(u_t) \tag{6}$$

These probabilities are used to compute the artifact-weighted loss defined in Equation 5, where each token's contribution is scaled by its corresponding importance score.

3.4 Dataset Curation via Token Diagnostics

After the classification task, to evaluate Artifact-Guided pretraining, we analyze token-level reconstruction loss (Figure 1(c)). Our assumption is to have the possibility to identify statistically important tokens that are difficult-to-reconstruct at MLM training stage as they can reflect distributional noise or semantic ambiguities that impair downstream performance. To that extent, we introduce a token scoring function based on label noise estimation, that aims to quantify token relationship with noise.

Noise-Driven Token (ND) Score We introduce a token-level scoring scheme based on annotation uncertainty. Following principles from confident learning (Northcutt et al., 2022), we identify frequent tokens in samples flagged as likely mislabeled, namely those having high discrepancies between predicted labels and class-conditional noise expectations. We extract these candidates using the `Cleanlab` toolkit (Northcutt et al., 2022), employing a confusion score that quantifies semantically contextually ambiguous tokens with the following score :

$$S_t = \frac{f_t}{\max_{t'} f_{t'}} \tag{7}$$

where f_t is the frequency of token t in potentially misannotated instances, normalized by the maximum token frequency among all such instances. To focus on informative textual features, we exclude stop words and retain only the top 25% most frequent tokens within this error subset.

Noise Removal Algorithm For each masked token, we compute reconstruction loss during pre-training and pair it with the computed importance score, aiming to curate dataset label by flagging text instances that contain high-scoring tokens.

Such a methodology allows human-in-the-loop intervention and exploratory analysis of candidate tokens that the user can iterate in (ranking) order to investigate and curate label noise prior to perform SUD classification model fine-tuning (Figure 1(c)).

Our intuition relies on the fact that frequent tokens likely belong to patterns recognized and learned by the model to generalize over SUD classes. In this manner, we propose to consider such token-level statistics across the relative model reconstruction capabilities at pretraining stage. To correct noisy instances in a given corpus, we propose two strategies:

(1) **Relabeling** that replaces a label of an instance X, if this latter is different from the instance label in which the token with the highest score (in X) occurs globally more often.

Example 1 Given a corpus $\mathcal{C}$ and a text instances $X = \{"abc"\} \in \mathcal{C}$ assigned with label 0. Let us consider that "b" would be the token with the highest score that occurs more often in instances assigned with label $1 \in \mathcal{C}$, the label of X changes to 1.

(2) **Filtering**, that removes from the corpus an instance X if this latter is assigned with a label different from the instance label in which the token with the highest score (in X) occurs globally more often.

Example 2 Given a corpus $\mathcal{C}'$ and a text instances $Y = \{"cde"\} \in \mathcal{C}'$ assigned with label 1. Let us consider that "d" would be the token with the highest score that occurs more often in instances assigned with label $0 \in \mathcal{C}'$, Y is removed from $\mathcal{C}'$.

Downstream Task For the evaluation of our methods we fine-tune our models on a multi-class classification task targeting Socially Unacceptable Discourse (SUD), as defined in Section 3.1.

The model architecture comprises a pretrained encoder that produces contextualized token representations, which are aggregated and passed to a lightweight classification head, a linear projection followed by a softmax activation, to yield a probability distribution over the target classes. Training is conducted using the standard cross-entropy loss between predicted distributions and ground-truth labels (Devlin et al., 2019; Clark et al., 2020; Zhang et al., 2023; Yang et al., 2023a; Moon et al., 2020; Sun et al., 2019).

4 Experimental Setup

In this section, we present the experimental framework employed to evaluate our artifact-aware pretraining approach, detailing the datasets and model configurations. Model training and evaluation were carried out using key libraries such as `transformers`, `datasets`, `PyTorch`, and `TensorFlow`. Comprehensive information on package versions and the computational environment is available in our repository to facilitate reproducibility: `https://github.com/rayaneghilene/MLM_Pretraining`.

4.1 Datasets

We utilize the G^{SUD} dataset introduced by Carneiro et al. (2023), which aggregates 13 publicly available English-language datasets spanning up to 12 SUD classes (see Table 4 in the appendix). The full corpus comprises approximately 500K instances, with a significant imbalance across classes. The *neither* class accounts for over 70% of the samples, while individual SUD categories vary in frequency. By combining sources that differ in both discourse and annotation practices, G^{SUD} captures substantial variation in the definition and expression of categories across contexts, providing a demanding benchmark for multi-domain learning. All datasets are publicly available and released under permissive licenses. Our use of these datasets is consistent with their original purpose, which was primarily classification and moderation of hate speech-related content. More details on the G^{SUD} dataset are found in the work of Carneiro et al. (2023).

4.2 Models and Training Setup

We experiment with four transformer-based models from the Hugging Face library: `bert-base-uncased` (110M parameters), `bert-large-uncased` (340M), `roberta-base` (125M), and `roberta-large` (355M). All experiments were conducted on an infrastructure equipped with NVIDIA A100 GPUs (80 GB memory) and 2 TB main memory, equipped with 2 AMD Milan EPYC 7543 processors (32 cores at 2.80 GHz).

MLM Pretraining Pretraining is conducted on the G^{SUD} corpus ($\sim$155K samples), to which we apply stratified under-sampling of the dominant *neither* class (10% retention).[3] Hyperparameters are tuned empirically. The final setup includes a learning rate of 1×10^{-5}, weight decay of 0.001, batch size of 128, and 5 training epochs.[4]

Downstream Classification Task Following pretraining, models are fine-tuned on each single dataset composing G^{SUD} for multi-class text classification. We split data using the following ratio: 80% training, 10% validation, and 10% test using stratified sampling. We employ the Hugging Face `AutoModelForSequenceClassification` wrapper, to attach a fully connected classification head

[3]A secondary configuration was also evaluated consisting of a focused subset containing only SUD-labeled instances ($\sim$120K samples), emphasizing domain-specific language. However, as this configuration did not yield any noticeable differences we omit to report the relative results.

[4]The hyperparameter search space included the following configurations: learning rate $\in \{5 \times 10^{-5}, 2 \times 10^{-5}, 1 \times 10^{-5}\}$, weight decay $\in \{0.1, 0.01, 0.001\}$, batch size $\in \{64, 128\}$, and number of epochs $\in \{3, 5\}$.

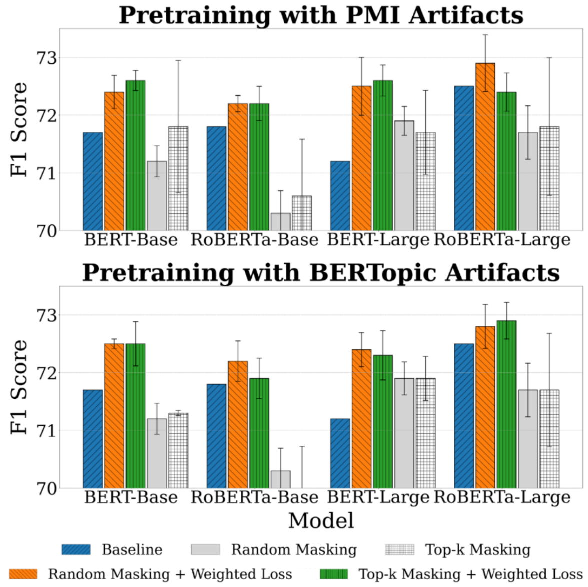

Figure 3: Aggregated F1 Score Performance Across Pretraining Paradigms and Experimental Conditions for BERT-Base, RoBerta-Base, BERT-Large, and RoBERTa-Large.

to the pretrained encoder. Input sequences are tokenized using the corresponding `AutoTokenizer` for each model. Fine-tuning is conducted for three epochs with a batch size of 16, a learning rate of 2×10^{-5}, and weight decay of 0.01. Each model is fine-tuned and evaluated across 10 runs with different seeds. Model performance is reported as the mean and standard deviation of the macro-averaged F1-score on the held-out test set.

5 Results

In this section we present and discuss the results of the proposed SUD classification framework.

5.1 Artifact-based Pretraining

Pretraining Strategy Evaluation Our first research objective is to evaluate the effectiveness of the different pretraining strategies discussed and proposed in the paper. Figure 3 shows the mean F1 score of SUD classification conducted with the baseline models and different pretraining strategies (**Random Masking + Weighted Loss**, **Top-k Masking + Weighted Loss**, **Random Masking**, **Top-k Masking**) leveraging the proposed artifacts scores (PMI Figure 3(top) and BERTopic Figure 3(bottom)). Here, we report the global mean of the classification performed in all the dataset reported in Table 4 across all masking proportions (Top-k = 5%, 10%, 15%, 25%, 35%). Our results indicate that pretraining strategies incorporating Weighted Loss exhibit the best performance across all the settings, confirming the hypothesis that models reinforce their generalizability when reconstructed loss is weighted according to the context importance.

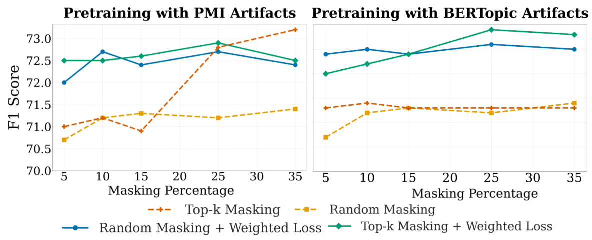

Figure 4: F1 Score Performance of BERT-Base Across Different Masking Percentages

Masking Percentage Optimization Figure 4 presents the F1 scores of BERT-Base across different masking percentages under Pretraining with PMI and BERTopic artifact masking. Again, weighted Loss strategies consistently outperform their non-weighted counterparts, yielding optimal performances at the 25% masking level. In general we note that BERT-Base has performance either on par, or superior than the other models. Hence, in the following part of the evaluation, we solely consider BERT-Base.

5.2 Dataset Curation via Token Diagnostics

To analyze the errors and individuate the effectiveness of artifact-guided pretraining in learning meaningful token representations, we analyse the correlation between token-level statistical score and reconstruction difficulty (model loss) in the whole G^{SUD} using the BERT-Base model. Figure 5 visualizes mean token reconstruction loss as a function of artifact extraction score (log scale), limited to the top 25% of tokens per method. Each dot represents a single token. To quantify these relationships, we compute Pearson correlation (r) between reconstruction loss and the log-transformed artifact scores. For the Noise-Driven score, we observe a strong negative correlation ($r = -0.64$), indicating that tokens with higher importance scores tend to be reconstructed more accurately in our framework. This suggests that tokens frequently associated with noisy labels are well reconstructed by the MLM reflecting the model's ability to encode patterns that are informative for reconstructing these artifacts. In contrast, BERTopic ($r = 0.01$) and PMI ($r = 0.03$) show no correlations with reconstruction loss, indicating a weaker association between

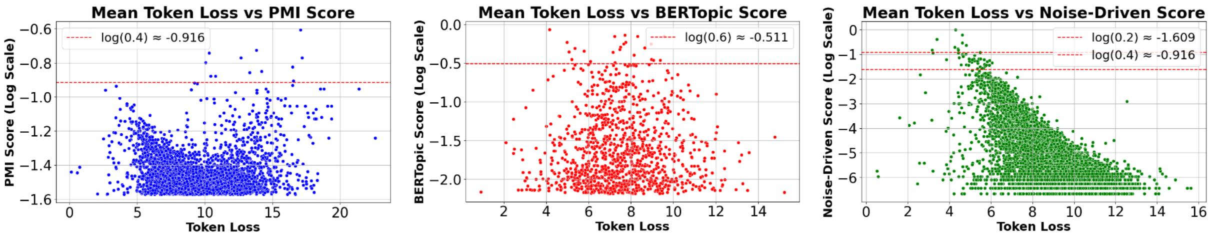

Figure 5: Mean token reconstruction loss vs. artifact extraction score (log scale). Each dot represents a single artifact from the top 25% of scores.

their frequency and model uncertainty during pre-training.

Dataset-Specific Curation We begin by evaluating our curation strategy within each dataset independently, leveraging the proposed scores vs. loss diagnostics. Specifically, we curate our dataset by flagging sentences containing high-scoring tokens selected by the following thresholds (depicted in Figure 5): **PMI** ≥ 0.4, **BERTopic** ≥ 0.6, and **Noise-Driven Score** ≥ 0.2 or ≥ 0.4. These values were selected to retain only the most salient tokens based on distributional breakpoints and qualitative review. Thresholds are reported here in raw form, while log-transformed values are used for visualization in plots. We then apply the token-level diagnostics described in Section 3.4 to curate the datasets with the two types of intervention: (1) **Relabeling**, and (2) **Filtering** prior to perform SUD classification.

Table 1 presents the performance impact of each method across 13 datasets. We observe that curation strategies yield mixed results: in datasets such as **Founta**, **Gab**, and **Hateval**, both relabeling and filtering significantly improve performance, suggesting that model errors were at least partly driven by annotation inconsistencies or label noise. Conversely, for datasets such as **Davidson** and **Olid**, the gains are modest or neutral, and aggressive filtering thresholds can even lead to degradation. Importantly, we found that certain datasets exhibited unexpected performance drops under these interventions, particularly **Fox** and **Grimminger**.

Large Scale Curation We extend our analysis considering SUD classification in the complete (G^{SUD}) corpus. Such scenario is challenging, as it introduces further annotation noise due to heterogeneous labeling criteria of different sources. We apply the same token-level relabeling and filtering methods described in the previous experiment. Table 2 (left) shows the absolute F1 scores for each

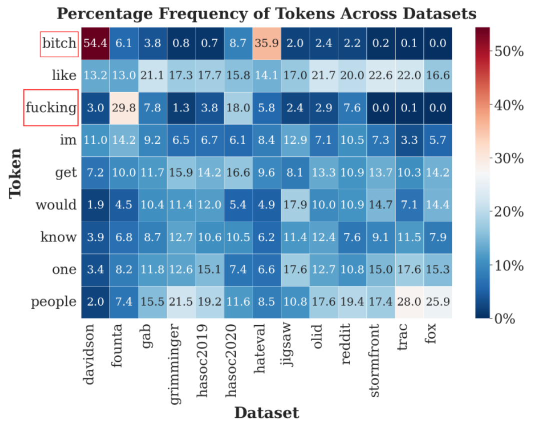

Figure 6: Percentage Frequency of Selected Tokens by Noise-Driven Method across Datasets

method on the G^{SUD} corpus ($\sim$155K samples) (we apply stratified under-sampling of the dominant *neither* class (10% retention) as suggested in (Carneiro et al., 2023)). Relabeling with noise-driven diagnostics at a 0.2 threshold performs best, achieving an F1 score of 66.9. Table 2 (right) reports the relative differences from the baseline. The same method yields the highest gain (+11.2), followed by BERTopic (+8.1). Filtering methods show smaller but consistent improvements. These results indicate that our token-based techniques are effective not only for individual datasets but also when applied to a large scale scenario, better resolving cross-dataset inconsistencies.

We also conducted a manual verification SUD Instance relabelling in the Appendix

5.3 Human-Guided Explainable Approach

While applying automatic relabeling and filtering strategies by automatically filtering noisy instances, we observe that even limited interventions could lead to notable drops in performance on certain datasets, such as **Grimminger** and **Fox** as shown in Table 1. To understand this, we analyze token distributions depicted in Figure 6, which displays the normalized frequency of selected tokens in each

Dataset	Baseline	Relabeling				Filtering			
		Noise-Driven $\geq$ 0.2	Noise-Driven $\geq$ 0.4	BERTopic $\geq$ 0.6	PMI $\geq$ 0.4	Noise-Driven $\geq$ 0.2	Noise-Driven $\geq$ 0.4	BERTopic $\geq$ 0.6	PMI $\geq$ 0.4
Davidson	$76.3_{1.2}$	$73.3_{2.2}$	$75.8_{2.3}$	$75.5_{1.6}$	$76.3_{1.2}$	$72.2_{3.7}$	$75.8_{1.5}$	**76.5**$_{1.7}$	$76.3_{1.2}$
Founta	$76.3_{0.7}$	**79.9**$_{1.5}$	**78.6**$_{0.9}$	**76.7**$_{0.6}$	$76.3_{0.7}$	**79.6**$_{1.6}$	**78.4**$_{0.9}$	$75.9_{0.7}$	$76.3_{0.7}$
Fox	$67.0_{2.7}$	$60.2_{4.6}$	$58.7_{3.1}$	$64.5_{0.7}$	$66.8_{2.7}$	$55.6_{6.3}$	$62.3_{4.2}$	$64.5_{4.3}$	$66.9_{2.7}$
Gab	$90.0_{0.3}$	**95.2**$_{0.3}$	**93.2**$_{0.5}$	**90.4**$_{0.5}$	**90.3**$_{0.4}$	**93.1**$_{0.5}$	**91.8**$_{0.5}$	**90.4**$_{0.6}$	$90.0_{0.3}$
Grimminger	$72.5_{2.5}$	$52.7_{4.5}$	$67.8_{5.8}$	$70.1_{0.3}$	$72.2_{2.2}$	$58.7_{3.7}$	$68.9_{5.5}$	$72.1_{3.6}$	$72.4_{2.5}$
Hasoc2019	$46.7_{1.8}$	$45.2_{3.4}$	$42.1_{2.4}$	$44.1_{3.3}$	$46.7_{1.8}$	$43.5_{3.8}$	$44.1_{2.9}$	$46.0_{1.3}$	$46.7_{1.8}$
Hasoc2020	$56.4_{2.2}$	$53.1_{3.9}$	$54.3_{2.3}$	$54.0_{2.0}$	$53.4_{2.2}$	**58.1**$_{3.6}$	$52.8_{2.2}$	**56.7**$_{2.9}$	$56.4_{2.2}$
Hateval	$77.0_{1.0}$	**87.1**$_{0.7}$	**84.5**$_{0.7}$	**77.7**$_{1.3}$	$77.0_{1.0}$	**84.9**$_{1.8}$	**80.9**$_{1.2}$	**77.3**$_{0.5}$	$77.0_{1.0}$
Jigsaw	$55.1_{1.0}$	**63.6**$_{2.0}$	**61.3**$_{1.1}$	$57.9_{0.9}$	$54.6_{1.2}$	**60.1**$_{2.7}$	**58.2**$_{1.2}$	**55.4**$_{0.9}$	$54.7_{1.0}$
Olid	$78.0_{1.0}$	**79.7**$_{1.5}$	**78.8**$_{1.0}$	$77.3_{1.1}$	$77.4_{1.0}$	**80.3**$_{2.2}$	$77.7_{1.3}$	$77.7_{3.6}$	$77.5_{1.0}$
Reddit	$84.8_{0.9}$	$82.5_{1.5}$	**86.5**$_{1.2}$	$84.5_{0.6}$	**84.9**$_{0.7}$	$83.7_{1.8}$	$86.4_{1.3}$	**85.1**$_{0.8}$	**84.9**$_{0.8}$
Stormfront	$77.9_{1.7}$	$63.6_{4.2}$	$70.9_{2.0}$	$75.9_{0.2}$	$77.8_{1.8}$	$69.1_{3.7}$	$75.2_{2.5}$	**78.1**$_{2.2}$	$77.9_{1.7}$
Trac	$74.6_{0.6}$	**80.4**$_{1.4}$	**78.2**$_{1.2}$	**74.9**$_{1.3}$	**74.7**$_{0.5}$	**79.1**$_{0.9}$	**77.5**$_{1.3}$	**75.0**$_{1.4}$	$74.1_{1.3}$

Table 1: Performance comparison of our relabeling and filtering methods at a dataset level. Scores are averages of 10 runs with different seeds, while subscripts indicate standard deviation. We depict scores above the baseline in bold.

Method	Relabeling	Filtering
Noise-Driven ($\geq$ 0.2)	$66.9_{1.6}$	$62.2_{1.4}$
Noise-Driven ($\geq$ 0.4)	$62.8_{1.1}$	$61.4_{1.3}$
BERTopic ($\geq$ 0.6)	$65.1_{0.5}$	$60.7_{0.6}$
PMI ($\geq$ 0.4)	$60.1_{0.7}$	$60.2_{0.4}$

Table 2.1: F1 scores for relabeling and filtering methods.

Method	Relabeling	Filtering
Noise-Driven ($\geq$ 0.2)	**+11.1**	**+3.3**
Noise-Driven ($\geq$ 0.4)	**+4.3**	**+2.0**
BERTopic ($\geq$ 0.6)	**+8.1**	**+0.8**
PMI ($\geq$ 0.4)	-0.2	+0.0

Table 2.2: Relative differences from baseline.

Table 2: Comparison of relabeling and filtering methods (left) on a class-balanced subset of the G^{SUD} corpus in F1 scores and their relative differences from baseline (right). Scores are averages of 10 runs with different seeds, while subscripts indicate standard deviation.

Dataset	Baseline	Relabeling			Filtering		
		ND $\geq$ 0.2	ND $\geq$ 0.4	BT $\geq$ 0.6	ND $\geq$ 0.2	ND$\geq$ 0.4	BT $\geq$ 0.6
Davidson	$76.3_{1.2}$	$75.5_{1.4}$	**77.3**$_{2.1}$	**76.5**$_{1.0}$	$76.2_{1.7}$	**76.6**$_{1.8}$	**76.5**$_{1.0}$
Founta	$76.3_{0.7}$	**78.4**$_{0.5}$	**78.2**$_{0.7}$	**78.1**$_{0.7}$	**78.2**$_{1.0}$	**77.8**$_{0.5}$	**77.9**$_{0.7}$
Fox	$67.0_{2.7}$	**67.2**$_{4.1}$	**70.6**$_{3.6}$	**70.3**$_{3.4}$	$65.7_{5.2}$	**70.4**$_{3.6}$	**70.6**$_{3.6}$
Gab	$90.0_{0.3}$	**92.4**$_{0.3}$	**91.0**$_{0.4}$	**90.8**$_{0.4}$	**92.2**$_{0.4}$	**91.0**$_{0.5}$	**90.8**$_{0.4}$
Grimminger	$72.5_{2.5}$	$68.4_{3.3}$	**72.7**$_{3.2}$	$72.5_{3.0}$	$71.4_{5.2}$	**72.6**$_{3.1}$	**72.6**$_{3.1}$
Hasoc2019	$46.7_{1.8}$	**48.5**$_{2.5}$	$46.4_{2.0}$	$46.0_{2.0}$	**47.8**$_{2.4}$	$45.6_{2.0}$	$46.0_{2.0}$
Hasoc2020	$56.4_{2.2}$	**58.2**$_{4.3}$	$56.2_{1.7}$	$56.2_{1.7}$	**56.7**$_{2.7}$	$56.2_{1.7}$	$56.2_{1.7}$
Hateval	$77.0_{1.0}$	**82.6**$_{0.9}$	**82.2**$_{0.9}$	**81.8**$_{0.9}$	**83.3**$_{0.7}$	**82.5**$_{0.9}$	**81.8**$_{0.9}$
Jigsaw	$55.1_{1.0}$	**64.5**$_{1.5}$	**58.2**$_{0.9}$	**58.4**$_{0.9}$	**60.5**$_{0.8}$	**58.7**$_{1.1}$	**58.4**$_{0.9}$
Olid	$78.0_{1.0}$	**80.1**$_{0.8}$	$77.7_{0.7}$	$77.6_{0.7}$	**78.6**$_{1.1}$	$77.7_{0.7}$	$77.6_{0.7}$
Reddit	$84.8_{0.9}$	**86.7**$_{1.0}$	**85.7**$_{0.9}$	**85.9**$_{0.7}$	**86.4**$_{0.8}$	**85.1**$_{1.1}$	**85.9**$_{0.7}$
Stormfront	$77.9_{1.7}$	$76.3_{2.5}$	**78.1**$_{1.8}$	**78.1**$_{1.8}$	$77.8_{2.1}$	**78.1**$_{1.1}$	**78.1**$_{1.8}$
Trac	$74.6_{0.6}$	**75.2**$_{1.6}$	**75.1**$_{0.7}$	**75.1**$_{0.7}$	**75.4**$_{1.2}$	**75.2**$_{0.6}$	**75.2**$_{0.7}$

Table 3: Comparison of F1 scores (mean_{std}) across datasets with different noise-handling strategies: baseline (original labels), relabeling using Noise-Driven (ND) and BERTopic (BT) approaches, and filtering based on noise scores.

dataset, for which the ND score is greater than or equal to 0.4. We observe that salient tokens potentially related to SUD (e.g., *b*tch*, *f*cking*) are under-represented in **Grimminger** and **Fox**, suggesting that previously repaired instances contain neutral language. Guided by such explanation, we restrict the interventions to tokens with high support and semantic relevance to SUD. In this experiment, we also consider token ranking using BERTopic-based score. This selective approach leads to more consistent improvements in ten datasets as reflected in Table 3.

We also perform a manual verification of SUD instance relabelling (See section 7 in the appendix).

6 Conclusion

The results of this study offer several insights into the role of SUD artifacts in guiding pretraining and improving dataset quality for the task of SUD classification. While standard masked language modeling provides only limited improvements in downstream performance across model architectures, introducing artifact-weighted loss consistently yields better results. By assigning greater importance to semantically important tokens, the model is encouraged to focus on contextually challenging regions. Beyond model performance, our curation strategy, based on token-level relabeling and filtering, proves valuable for interpretability. With minimal changes to the data, the method reveals annotation inconsistencies, offering a lightweight mechanism for surfacing potential labeling issues. This enables more transparent error analysis and targeted refinement.

7 Limitations

While our artifact-guided curation framework demonstrates clear potential, it can further evolve to better support human understanding during annotation. Although our approach effectively surfaces tokens in ambiguous or noisy environments through artifact-based heuristics, it may overlook subtle linguistic inconsistencies or contextual errors that can only be reliably detected by human annotators. As a result, a dedicated annotation campaign supported by our solution remains necessary to validate and complement our dataset curation methods. Moreover, we plan to extend the human-in-the-loop approach to the G^{SUD} corpus, increasing our ability to assess the method's effectiveness in this setting, where tailored adaptation and computational challenges need to be addressed. Finally, the token selection thresholds, though informed by distributional patterns and qualitative assessment, remain heuristic. Future work could investigate more principled, data-driven approaches to enhance the robustness and generalizability of the dataset curation framework.

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

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, et al. 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*.

Bruno Machado Carneiro, Michele Linardi, and Julien Longhi. 2023. Studying socially unacceptable discourse classification (SUD) through different eyes: "are we on the same page ?". *CoRR*, abs/2308.04180.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Thomas Davidson, Dana Warmsley, Michael W. Macy, et al. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM*. AAAI Press.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Koustuv de Maiti and Darja Fišer. 2021. Working with socially unacceptable discourse online: Researchers' perspective on distressing data. In *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)*, pages 78–82.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*.

Danielle Draper and Sabine Neschke. 2023. Social media algorithms: The pros and cons.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert M Fano. 1963. *Transmission of information: a statistical theory of communications*. MIT press.

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

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, et al. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM*.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*.

Rayane Ghilene, Dimitra Niaouri, Michele Linardi, and Julien Longhi. 2024. Analysis of socially unacceptable discourse with zero-shot learning. *Preprint*, arXiv:2409.13735.

Lara Grimminger and 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. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EACL 2021*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3309–3326. Association for Computational Linguistics.

Thorsten Joachims et al. 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML*, volume 97, pages 143–151. Citeseer.

Vanja Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Jan Kocon, Alicja Figas, Marcin Gruza, et al. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Inf. Process. Manag.*, 58(5).

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, et al. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2020. Pmi-masking: Principled masking of correlated spans. *Preprint*, arXiv:2010.01825.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*.

Thomas Mandl, Sandip Modha, Anand Kumar M, et al. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, et al. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*.

Ilia Markov and Walter Daelemans. 2021a. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22.

Ilia Markov and Walter Daelemans. 2021b. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection. *Preprint*, arXiv:2012.10289.

Yu Meng, Jitin Krishnan, Sinong Wang, Qifan Wang, Yuning Mao, Han Fang, Marjan Ghazvininejad, Jiawei Han, and Luke Zettlemoyer. 2024. Representation deficiency in masked language modeling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. 2020. Masker: Masked keyword regularization for reliable text classification. In *AAAI Conference on Artificial Intelligence*.

Nafeesa Murad, Mohd Hilmi Hasan, Muhammad Azam, Nadia Yousuf, and Jameel Yalli. 2024. Unraveling the Black Box: A Review of Explainable Deep Learning Healthcare Techniques. *IEEE Access*, PP:1–1.

Dimitra Niaouri, Michele Linardi, and Julien Longhi. 2024. Towards a new contextualized annotation schema for unacceptable and extreme speech (cues) to unleash generalization capability of ml models. *Studii de lingvistică*, 14(2):63–94.

Dimitra Niaouri, Bruno Machado Carneiro, Michele Linardi, and Julien Longhi. 2025. 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? *Digital linguistics*.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2022. Confident learning: Estimating uncertainty in dataset labels. *Preprint*, arXiv:1911.00068.

Inez Okulska and Anna Kołos. 2024. A morpho-syntactic analysis of human-moderated hate speech samples from wykop.pl web service. *Półrocznik Językoznawczy Tertium*, 8:54–71.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics: Student Research Workshop, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Joni Salminen, Maximilian Hopf, Shammur Chowdhury, Soon-Gyo Jung, Hind Almerekhi, and Jim Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. *Preprint*, arXiv:1911.03891.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826, Seattle, United States. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *ArXiv*, abs/1904.09223.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019a. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 940–950.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019b. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Betty Van Aken, Julian Risch, Ralf Krestel, et al. 2018. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*.

Vasja Vehovar, Blaž Povž, Darja Fiser, Nikola Ljubešić, Ajda Šulc, and Dejan Jontes. 2020. Družbeno nesprejemljivi diskurz na facebookovih straneh novičarskih portalov. *Teorija in Praksa*, 57:622–645.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-BART: Leveraging toxicity attributes for explanation generation of implicit hate speech. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13967–13983, Bangkok, Thailand. Association for Computational Linguistics.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2023a. Learning better masking for better language model pre-training. *Preprint*, arXiv:2208.10806.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023b. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505. Association for Computational Linguistics.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander F Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification. In *IberLEF@ SEPLN*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Lanqin Yuan and Marian-Andrei Rizoiu. 2022. Detect hate speech in unseen domains using multi-task learning: A case study of political public figures. *CoRR*, abs/2208.10598.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, et al. 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*.

Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, Xin Cao, Kongzhang Hao, Yuxin Jiang, and Wei Wang. 2023. Weighted sampling for masked language modeling. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Appendix

Datasets

In Table 4, we report the source and metadata of all datasets used in the empirical evaluation of our solution.

Dataset	Source	Sample Type	# Samples	Labels
Davidson	Davidson et al. (2017)	Tweets	25,000	hate, offensive, neither
Founta	Founta et al. (2018)	Tweets	100,000	abusive, hate, neither
Fox	Gao and Huang (2017)	Threads	1,528	hate, neither
Gab	Kocon et al. (2021)	Posts	34,000	hate, neither
Grimminger	Grimminger and Klinger (2021)	Tweets	3,000	hate, neither
HASOC2019	Mandl et al. (2019)	Facebook, Twitter	12,000	hate, offensive, profane, neither
HASOC2020	Mandl et al. (2020)	Facebook posts	12,000	hate, offensive, profane, neither
Hateval	Basile et al. (2019)	Tweets	13,000	hate, neither
Jigsaw	Van Aken et al. (2018)	Wikipedia talk pages	220,000	identity hate, insult, obscene, severe toxic, threat, toxic, neither
Olid	Zampieri et al. (2019)	Tweets	14,000	offensive, neither
Reddit	Yuan and Rizoiu (2022)	Posts	22,000	hate, neither
Stormfront	De Gibert et al. (2018)	Threads	10,500	hate, neither
Trac	Kumar et al. (2018)	Facebook posts	15,000	aggressive, neither

Table 4: Summary of datasets used in this study Carneiro et al. (2023).

Manual Verification of SUD Instance Relabelling

To qualitatively evaluate the proposed Dataset Curation method (Section 5.2), we conducted a manual review of a representative sample. In Tables 5 and 6 we present the manual verification of instances identified by our token-driven diagnostic algorithms (Noise-Driven and BERTopic Scores). Each row lists the sentence (Instance), the salient Token triggering the flag, the original Ground Truth label, the suggested Relabeling, the Dataset source, and the outcome of Manual verification.

For this inspection, we applied a token-driven extraction strategy, selecting sentences linked to the most salient tokens (Noise-Driven score ≥ 0.4, BERTopic score ≥ 0.6). Our findings indicate that the framework reliably identifies annotation inconsistencies and borderline cases that benefit from closer human scrutiny.

In many instances, the method identifies clear annotation errors that warrant relabeling:

- Table 5 (18): *"and this retarded bitch has over k followers"* (Neither → Hate)

In other cases, relabeling is suggested where the presence of vulgar language makes the distinction between socially unacceptable and acceptable content ambiguous, requiring human judgment:

- Table 6 (2): *"fuck all that bullshit who tf is young crazy ep dropping this april stay tuned"* (Neither → Abusive)

Occasionally, the algorithm produces incorrect relabelings showing that even those cases allow us to identify spurious correlations, where tokens typically associated with SUD may in fact occur in a simple vulgar context:

- Table 5 (5): *"frustration can be a bitch"* (Neither → Offensive)

Overall, these results demonstrate that our method can serve as an effective diagnostic tool, complementing human judgment and aiding annotators in refining datasets with improved precision.

Instance	Token	Ground Truth	Relabeling	Dataset	Manual verification
1. *all of kate uptons leaked nudes today amp more bitch lo*	*bitch*	Neither	Offensive	Davidson	Correct
2. *mouth fucking for two bitches*	*fucking*	Neither	Abusive	Founta	Correct
3. *my favorite episode of friends is the one where blacks mysteriously vanish and unemployed white people find affordable ho*	*people*	Neither	Offensive	Davidson	Correct
4. *it is u and sickular mind people*	*people*	Neither	Aggressive	Trac	Correct
5. *frustration can be a bitch*	*bitch*	Neither	Offensive	Davidson	Incorrect
6. *you do and so does sb its about forced deviance without that no sense would be possible*	*would*	Hate	Neither	Fox	Correct
7. *so and his wife get a payout that was intended for small business but individuals dont get monetary relief*	*get*	Hate	Neither	Grimminger	Correct
8. *that would suck*	*would*	Profane	Neither	Hasoc2020	Correct
9. *look like merkel just want to get laid*	*like*	Hate	Neither	Fox	Incorrect
10. *im not doing this goddamn*	*im*	Insult	Neither	Jigsaw	Correct
11. *we know the real too and america cant survive four more years*	*know*	Offensive	Neither	Hasoc2019	Correct
12. *whoever the last bitch you was fucking with tell her its over*	*fucking*	Neither	Hate	Hateval	Correct
13. *he is hu and no one has checked on him*	*one*	Offensive	Neither	Olid	Correct
14. *if u acted like a hoe after we broke up im not wrong for thinking u were a hoe all along*	*like*	Neither	Offensive	Davidson	Correct
15. *im with you too what*	*im*	Hate	Neither	Reddit	Correct
16. *hell to have a teacher like this*	*like*	Hate	Neither	Stormfront	Correct
17. *laughing my fucking ass off beechhouse of cards is collapsing*	*fucking*	Neither	Hate	Grimminger	Incorrect
18. *and this retarded bitch has over k followers*	*bitch*	Neither	Hate	Gab	Correct
19. *im ready are you vs*	*im*	Hate	Neither	Grimminger	Correct
20. *i m not sure you should assume im not corrupt joe biden*	*im*	Hate	Neither	Grimminger	Correct
21. *this bitch thincc she dress coding me*	*bitch*	Neither	Offensive	Davidson	Correct

Table 5: Manual verification of instances flagged by our token-driven diagnostic algorithm (Noise-Driven Score). Columns show the sentence (**Instance**), salient **Token**, original **Ground Truth**, suggested **Relabeling**, **Dataset**, and **Manual verification** outcome (green = correct, red = incorrect).

Instance	Token	Ground Truth	Relabeling	Dataset	Manual verification
1. *miserable mahas pathetic pattys amp bullshit beckys*	*bullshit*	Neither	Abusive	Founta	Correct
2. *fuck all that bullshit who tf is young crazy ep dropping this april stay tuned*	*bullshit*	Neither	Abusive	Founta	Correct
3. *im so mad that it was allowed for him to be railroaded out on bullshit but brian williams is still out there*	*bullshit*	Neither	Abusive	Founta	Correct
4. *one pathetic piggy could barely hit before he creamed his pants try harder bitches this is just pathetic*	*piggy*	Neither	Abusive	Founta	Correct
5. *but yall make the mutual decision to have sex on the first date but shes trash what kind of math*	*math*	Neither	Offensive	Davidson	Correct
6. *leabian gets butt fucked and pussy eaten*	*butt*	Neither	Abusive	Founta	Correct
7. *the soviets managed to accomplish total usurpation criminalization radicalization and grabbed under kgbists total mob monopolization*	*criminal*	Hate	Neither	Founta	Correct
8. *all pakistani people are again in electronic indian journal its prove that they dont like their country rather merge with original country and stop killing both sides of army rather do on original criminal british people*	*criminal*	Neither	Aggressive	Trac	Correct
9. *by being a spoiled entitled ghetto head that really screams feminist*	*ghetto*	Neither	Offensive	Olid	Correct
10. *speech is like boring song where you go out and visit wash room*	*boring*	Neither	Aggressive	Trac	Incorrect
11. *as i commented a few days ago on this thread massive race or ethnic mixing tends to preceed massive race or ethnic war*	*thread*	Hate	Neither	Stormfront	Correct
12. *vip highdef bignipples gt gt gt*	*nipples*	Abusive	Neither	Founta	Incorrect
13. *r u fucking kidding me*	*kidding*	Neither	Abusive	Founta	Incorrect
14. *go eat some more burgers butthead*	*butthead*	Neither	Offensive	Olid	Correct
15. *nasty little butt slut nicole moore*	*butt*	Neither	Abusive	Founta	Correct
16. *niggas dont be savages they just be broke niggas with anger problems*	*niggas*	Neither	Hate	Founta	Correct
17. *what issssss that anger its not a conclusion to be right*	*anger*	Aggressive	Neither	Trac	Correct

Table 6: Manual verification of instances flagged by our token-driven diagnostic algorithm (BERTopic Score). Columns show the sentence (**Instance**), salient **Token**, original **Ground Truth**, suggested **Relabeling**, **Dataset**, and **Manual verification** outcome (green = correct, red = incorrect).