Robustness Evaluation of Hate Speech Detection Models Under Structured Adversarial Obfuscation

Anonymous ACL submission

Abstract

Hate speech classifiers are vulnerable to adversarial text obfuscation, yet existing robustness benchmarks typically evaluate models using either random noise or isolated perturbation techniques, failing to capture the structured multi-strategy evasion behaviour observed in practice. We present a systematic benchmark evaluating four transformer-based detectors against six obfuscation techniques under three attack regimes spanning *realistic*, *uniform*, and *adversarial* distributions. Beyond prior work, we additionally model realistic social media language by obfuscating neutral (non-toxic) tokens alongside toxic ones, and introduce *obfuscation intensity* as a dedicated evaluation axis. Our results show substantial robustness degradation across all models, with F1 drops of up to -0.356 and performance deteriorating non-linearly as perturbation density increases. Obfuscation-aware fine-tuning recovers up to $+0.326$ F1, demonstrating that robustness can be substantially improved without architectural modification.

1 Introduction

The proliferation of hate speech on social media platforms represents one of the most pressing challenges in online safety. A recent cross-national survey found that two in three people frequently encounter hate speech online (Ipsos, 2023), while platforms such as Facebook and Instagram removed over 14 million pieces of hateful content in a single quarter of 2024 alone (Meta Platforms, 2024). Beyond sheer volume, hate speech causes measurable psychological harm to targeted individuals and has been linked to real-world violence and discrimination (Vidgen and Derczynski, 2020). Automated detection systems have thus become an indispensable tool for large-scale content moderation.

The advent of transformer-based models has substantially advanced the state of the art in hate speech detection (Waseem and Hovy, 2016; Davidson et al., 2017; Caselli et al., 2021). Models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and domain-adapted variants like HateBERT (Caselli et al., 2021) achieve strong performance on standard benchmarks. However, these systems are evaluated almost exclusively on clean, unperturbed corpora — an assumption that rarely holds in adversarial real-world conditions.

In practice, users attempting to evade automated moderation employ *textual obfuscation*: deliberate manipulation of word surface forms to preserve semantic intent while bypassing classifiers. Strategies range from character substitution (*bitch* $\rightarrow$ *b!tch*) and phonetic replacement (*fool* $\rightarrow$ *phool*) to intra-word spacing (*moron* $\rightarrow$ *m o r o n*) and composite perturbations. Such evasion is well documented in the content moderation literature (Gröndahl et al., 2018), yet its systematic impact on transformer-based detectors remains critically understudied.

Existing work on obfuscation robustness suffers from two key limitations. Prior benchmarks either apply uniform random noise to all tokens — overestimating evasion impact — or target single isolated words, underestimating the structured multi-strategy behaviour of real adversaries (Aggarwal and Zesch, 2022). Moreover, evaluations treat obfuscation as binary, ignoring how perturbation *density* affects detection difficulty. We address both gaps with a comprehensive benchmark that models obfuscation as a probabilistic, policy-driven process spanning real-world, uniform, and adversarial attack distributions. We further introduce obfuscation intensity as a novel evaluation axis and construct a realistic dataset variant that obfuscates neutral tokens alongside toxic ones, faithfully simulating the informal register of social media users who casually abbreviate benign language while concealing harmful content.

Specifically, we examine: (1) the extent to which obfuscation degrades detection performance across

model architectures, (2) whether certain transformers exhibit greater robustness to adversarial perturbations, (3) how obfuscation intensity affects model performance, and (4) whether obfuscation-aware training can recover lost robustness.

Contributions.

- We present the **first systematic benchmark** evaluating four transformer models against six obfuscation techniques under three probabilistic policy regimes, with F1 drops of up to -0.356 under realistic evasion conditions.
- We introduce **obfuscation intensity** as a novel evaluation axis and a **realistic dataset variant** extending toxic-word perturbation to neutral tokens, exposing a non-linear degradation curve absent from prior work.
- We show that **obfuscation-aware fine-tuning** recovers up to $+0.326$ F1 with no architectural changes, providing a practical deployable defence for content moderation systems.

2 Related Work

2.1 Hate Speech Detection

Early approaches to hate speech detection relied on lexicon-based methods and traditional machine learning classifiers such as SVMs and logistic regression (Waseem and Hovy, 2016; Davidson et al., 2017). Davidson et al. (2017) introduced the widely adopted three-class Twitter corpus distinguishing hate speech, offensive language, and neither, which remains a standard benchmark. The introduction of transformer architectures (Devlin et al., 2019) marked a step change in detection performance. Caselli et al. (2021) demonstrated that domain-adaptive pre-training on abusive online content yields substantial gains over general-purpose encoders, motivating the inclusion of HateBERT in our benchmark. Fortuna and Nunes (2018) provided a comprehensive survey of hate speech detection methods, highlighting the persistent challenge of generalisation across datasets and linguistic registers — a challenge that obfuscation directly exacerbates.

2.2 Adversarial Robustness in NLP

The vulnerability of NLP models to adversarial perturbations has been extensively studied. Ebrahimi et al. (2018) introduced HotFlip, a white-box attack that exploits gradients to flip characters or tokens to change classification outcomes. Wallace et al. (2019) demonstrated that a fixed sequence of tokens, optimised by gradient search, could consistently reduce accuracy across many inputs, establishing the concept of universal adversarial triggers. Morris et al. (2020) unified this literature through TextAttack, a modular framework for adversarial attacks and data augmentation in NLP, providing standardised benchmarking infrastructure. Collectively, these works establish that surface-level perturbations — even character-level ones — can fundamentally undermine model robustness, a finding we extend specifically to the hate speech detection domain.

2.3 Obfuscation as Evasion of Hate Speech Detection

The intersection of adversarial NLP and hate speech moderation has received growing attention. Gröndahl et al. (2018) conducted a seminal study demonstrating that all proposed hate speech detection techniques are brittle against adversaries who insert typos, change word boundaries, or add innocuous words. Crucially, they showed that character-level features make models more attack-resistant than word-level features — an insight that informs our choice of subword tokeniser-based models. Röttger et al. (2021) introduced HateCheck, a suite of functional tests for hate speech detection models, revealing critical weaknesses in both academic and commercial systems. While HateCheck evaluates models on handcrafted test cases, it does not systematically vary obfuscation type or intensity across a realistic corpus, a gap our benchmark addresses directly.

Most closely related to our work, Aggarwal and Zesch (2022) analysed the real vulnerability of hate speech detection systems against targeted intentional noise, conducting a user study to identify which words users would actually obfuscate in a post. They found that real-world vulnerability is almost as high as white-box attacks, and much more severe than non-targeted dictionary methods. Our work differs in three important ways. First, rather than constraining obfuscation to a single token per post, we apply structured multi-technique perturbation under three probabilistic policy regimes, modelling a wider spectrum of adversarial behaviour. Second, we introduce obfuscation intensity as an explicit evaluation axis, quantifying how detection difficulty scales with perturbation density. Third, we extend evaluation beyond single-model settings

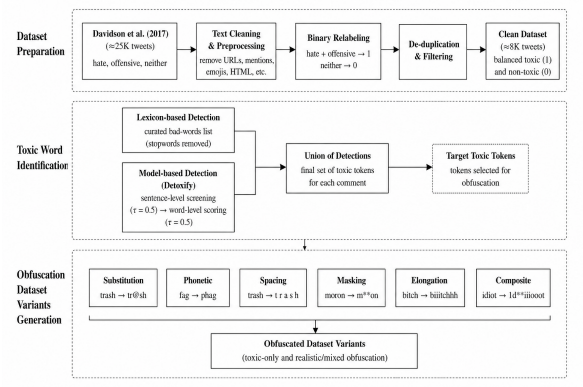

Figure 1: Overview of the obfuscation pipeline: dataset preparation (top), hybrid toxic word identification (middle), and obfuscation dataset variant generation (bottom).

to a controlled four-model benchmark, enabling cross-architecture robustness comparison that is absent from prior work. Existing work largely evaluates isolated perturbations or handcrafted attacks, but lacks a unified benchmark framework that jointly studies obfuscation type, perturbation density, and robustness recovery under controlled conditions.

3 Dataset and Obfuscation Framework

3.1 Source Dataset and Label Schema

We build upon the benchmark dataset introduced by Davidson et al. (2017), which comprises approximately 25,000 tweets annotated into three categories: *hate speech*, *offensive language*, and *neither*. For our binary classification objective, we consolidate the first two categories into a single positive class (label 1) representing toxic or hate speech, while retaining the third as the negative class (label 0). This relabelling is consistent with prior work that treats offensive and hate speech as a unified threat class when evaluating content moderation systems (Fortuna and Nunes, 2018; Founta et al., 2018).The overall dataset preparation and obfuscation generation pipeline is illustrated in Figure 1.

The final working corpus comprises 8,326 samples with a balanced class distribution of 4,163 toxic and 4,163 non-toxic instances, constructed through random sampling from the merged toxic class. This balanced setup ensures that accuracy and F1 are directly comparable across all experimental conditions without class-imbalance confounds.

3.2 Toxic Word Identification

A prerequisite for targeted obfuscation is accurate identification of toxic tokens within each utterance. Naively obfuscating all tokens would overestimate real-world evasion impact (Aggarwal and Zesch, 2022), while random single-token selection would underestimate it. We therefore employ a *hybrid detection strategy* that combines two complementary signals, as illustrated in the middle panel of Figure 1.

A lexicon-based filter matches surface forms against a curated bad-words list,[1] excluding common stopwords to suppress false positives; while effective for known slurs, it fails to capture context-dependent toxic language absent from the vocabulary (Fortuna and Nunes, 2018). To address this, we deploy *Detoxify* (Hanu and Unitary Team, 2020), which first screens the full utterance at sentence level ($\tau = 0.5$), then scores each candidate token individually (length > 3, non-stopword), retaining those exceeding τ — a threshold widely adopted in toxicity pipelines to balance precision and recall (Vidgen and Derczynski, 2020). The final toxic token set is the *union* of both detectors, ensuring high recall of context-dependent slurs that evade purely lexical methods while preserving coverage of well-known offensive terms that may fall below the model threshold in isolation.

3.3 Obfuscation Techniques

We implement six character- and token-level obfuscation operators, each simulating a distinct evasion strategy documented in real-world hate speech evasion (Gröndahl et al., 2018; Aggarwal and Zesch, 2022). The full pipeline — from dataset preparation through toxic word identification to obfuscation variant generation — is illustrated in Figure 1. Table 1 summarises each technique with a representative example.

Substitution applies a probabilistic character mapping (p = 0.6 per eligible character) using a leet-speak substitution table(e.g., $a \leftrightarrow @$, $e \leftrightarrow 3$, $i \leftrightarrow !$), guaranteeing at least one substitution per toxic token. **Phonetic** obfuscation replaces substrings with phonetically equivalent sequences (e.g., $ph \leftrightarrow f$, $oo \leftrightarrow u$) while protecting common digraphs (*ch*, *sh*, *th*) to preserve readability. **Spacing** inserts intra-word whitespace via single-split, double-split,

[1] Bad-words list sourced from Carnegie Mellon University: https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

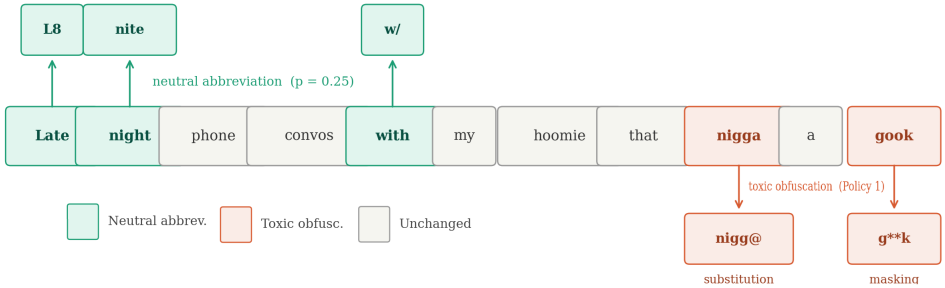

Figure 2: Illustration of the two-stage obfuscation pipeline on a representative comment from D_c. Neutral tokens are stochastically abbreviated upward (p = 0.25); toxic tokens are obfuscated downward via the Policy 1 selected technique (substitution: *nigga* → *nigg@*; masking: *gook* → *g**k*). Unchanged tokens are shown in grey.

Technique	Example
Substitution	trash → tr@sh
Phonetic	fool → phool
Spacing	hate → h a t e
Masking	moron → m**on
Elongation	bitch → biiitchhh
Composite	idiot → 1d**iiiooot

Table 1: Obfuscation techniques used for generating adversarial toxic text variants.

Dataset	Size	Description
D_a (Clean)	8,326	Original corpus; no perturbation
D_b (Mixed)	8,326	Toxic tokens obfuscated via 3-policy blend
D_c (Realistic)	8,326	D_b + neutral word abbreviation (p=0.25)
D_d (Individual)	8,326 each	One technique per variant (D1–D6)

Table 2: Dataset variants used for benchmarking.

or character-level decomposition, directly exploiting the token boundary assumptions of subword tokenisers such as BPE (Sennrich et al., 2016). **Masking** replaces a contiguous interior substring with special symbols ($*$, #), obscuring the word's core characters while preserving its overall shape. **Elongation** stochastically repeats characters 2–4 times at randomly selected positions, mimicking the emphatic spelling common in informal social media text (Eisenstein, 2013). Finally, **Composite** sequentially chains 2–3 randomly sampled techniques from the above set, generating the most adversarially challenging surface forms and most closely simulating sophisticated evasion behaviour observed in practice (Aggarwal and Zesch, 2022).

3.4 Dataset Variants

We construct four dataset variants (Table 2) to enable controlled benchmarking across a spectrum of obfuscation complexity. Figure 2 illustrates the two-stage obfuscation process on a representative comment.

D_a (Baseline) serves as the unperturbed reference, retaining the original cleaned text with no artificial perturbation. **D_b (Mixed Obfuscation)** introduces policy-level stochasticity: for each comment, one of three calibrated probability distributions (Table 3) is selected with equal probability (P = 0.33), and for each toxic token the technique is sampled from the chosen policy. **D_c (Realistic)** extends D_b by additionally applying an internet-abbreviation lexicon to non-toxic tokens (P = 0.25 per entry), simulating the informal register prevalent in social media and mitigating the risk that models learn to associate obfuscation patterns with toxicity rather than semantic content (Eisenstein, 2013; Tagliamonte and Denis, 2008). **D_d (Individual)** consists of six independent variants (D_1–D_6), each applying a single obfuscation technique exclusively to identified toxic tokens, enabling fine-grained attribution of model degradation to individual perturbation types.

3.5 Policy Distributions for Mixed Obfuscation

Three policies govern technique selection within D_b and D_c (Table 3). Unlike prior synthetic perturbation frameworks that apply uniformly random transformations (Gröndahl et al., 2018; Jones et al., 2020), our policies collectively span the space of plausible real-world evasion behaviours through non-uniform obfuscation priors.

P1 (Real-world) assigns higher probability to substitution (0.30) and phonetic alteration (0.20), consistent with prior observations that character-level manipulation is the most prevalent evasion strategy in online hate speech corpora (Gröndahl et al., 2018). **P2 (Uniform)** assigns equal probability (≈0.17) to all techniques, eliminating distributional bias and serving as a bias-free experimental

Policy	Subst.	Phonetic	Spacing	Masking	Elong.	Composite
P1 (Real-world)	0.30	0.20	0.10	0.05	0.15	0.20
P2 (Uniform)	0.17	0.17	0.17	0.17	0.17	0.17
P3 (Adversarial)	0.10	0.10	0.25	0.15	0.10	0.30

Table 3: Sampling distributions over obfuscation techniques for the three policy regimes used in mixed-obfuscation generation, modelling realistic, uniform, and adversarial evasion behaviour.

control for technique-level sensitivity analysis. **P3 (Adversarial)** up-weights composite (0.30) and spacing (0.25), modelling a worst-case attacker who preferentially selects transformations that fragment subword tokenisation boundaries — the primary structural assumption of transformer-based encoders (Ebrahimi et al., 2018; Wallace et al., 2019). For each comment, a single policy is drawn uniformly (P = 0.33) and applied consistently to all toxic tokens, preserving intra-comment stylistic coherence.

3.6 Neutral Word Obfuscation in D_c

To construct a maximally realistic evaluation set, D_c augments the mixed-obfuscation layer with stochastic substitution of neutral tokens using a curated lexicon of 80+ internet abbreviations, compiled from established internet language resources (Han and Baldwin, 2011; Eisenstein, 2013) and spanning four categories reflective of authentic social media register: chat shortcuts (*great→gr8*), internet slang (*oh my god→omg*), vowel dropping (*people→ppl*), and casual reductions (*kind of→kinda*). This categorisation is grounded in prior sociolinguistic work documenting the systematic nature of informal language on social media platforms (Tagliamonte and Denis, 2008).

Multi-word entries are matched with longest-first priority to prevent partial overlaps, and original capitalisation is preserved via case-aware replacement. Crucially, toxic tokens identified in §3.2 are explicitly protected from neutral obfuscation, preventing conflation of the two transformation layers. Each neutral token is substituted independently with probability $p = 0.25$, ensuring naturalness without over-abbreviation. As illustrated in Figure 2, the result is a comment where both harmful and benign tokens reflect realistic social media writing patterns simultaneously.

4 Experimental Setup

4.1 Models

We evaluate four transformer-based encoder models that collectively span the space of lightweight, high-capacity, domain-specialised, and multilingual architectures (Table 4). This selection is motivated by the goal of isolating whether obfuscation robustness is a function of model scale, pre-training data breadth, or domain specialisation — three axes identified as critical determinants of robustness in adversarial NLP (Morris et al., 2020; Wang et al., 2021).

Model	HuggingFace ID	Params	Pre-training
DistilBERT	`distilbert-base-uncased`	66M	General (EN)
RoBERTa	`roberta-base`	125M	General (EN)
HateBERT	`GroNLP/hateBERT`	110M	Hate speech
XLM-R	`xlm-roberta-base`	270M	Multilingual

Table 4: Pre-trained encoder models with HuggingFace identifiers, parameter counts, and pre-training corpora.

Table 4 summarises the four models. DistilBERT (Sanh et al., 2019) represents efficiency-constrained deployments; RoBERTa (Liu et al., 2019) establishes the general-purpose ceiling; HateBERT (**?**) tests whether domain specialisation confers robustness; and XLM-R (Conneau et al., 2020) evaluates cross-lingual pre-training breadth. All models are sourced from the HuggingFace Model Hub (Wolf et al., 2020) and fine-tuned under identical conditions.

We deliberately exclude autoregressive large language models such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023), as encoder-only architectures remain the dominant paradigm in deployed content moderation systems owing to their inference efficiency and direct suitability for sequence classification (Vidgen and Derczynski, 2020; Modha et al., 2020).

4.2 Experimental Conditions

We define six experimental conditions that progress from clean baseline evaluation through increasingly adversarial settings to a recovery analysis, summarised in Table 5.

CLEAN evaluates all models on unperturbed data, establishing the upper-bound performance ceiling under standard conditions. MIXED and REALISTIC transfer models trained on clean data to obfuscated test sets D_b and D_c respectively, directly

Condition	Train	Test	Objective
CLEAN	D_a	D_a	Baseline performance ceiling
MIXED	D_a	D_b	Robustness under mixed obfuscation
REALISTIC	D_a	D_c	Robustness under realistic obfuscation
PER-TECHNIQUE	D_a	D_d (×6)	Per-technique vulnerability analysis
INTENSITY	D_a	D_b (binned)	Effect of obfuscation density
RECOVERY	D_b	D_b	Obfuscation-aware training recovery

Table 5: Experimental conditions with train/test configurations and objectives.

measuring robustness degradation under controlled evasion. PER-TECHNIQUE disaggregates performance across each of the six individual obfuscation variants in D_d, enabling fine-grained attribution of degradation to specific perturbation types. INTENSITY partitions D_b by the number of obfuscated tokens per comment (0, 1, 2–3, $\geq$4), quantifying how detection difficulty scales with perturbation density — an evaluation axis absent from prior work. Finally, RECOVERY trains models directly on D_b and evaluates on D_b, testing whether obfuscation-aware fine-tuning can recover performance lost under adversarial conditions.

4.3 Training and Evaluation

All datasets are partitioned using a stratified 80/20 train–test split to preserve class balance across evaluation conditions. Full hyperparameter and implementation details are provided in Appendix A.

We adopt **F1** as the primary metric (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020), with full metrics (Accuracy, Precision, Recall, F1) reported for main conditions (Table 6) and F1 only for per-technique results (Table 7). Recall is additionally reported for intensity analysis to directly capture missed hate speech as perturbation density increases.

5 Results and Discussion

5.1 Baseline Performance and Obfuscation Impact

All models achieve strong F1 on clean data (D_a: 0.936–0.945), confirming that standard fine-tuning is sufficient for unperturbed hate speech detection. Under obfuscation (D_b), performance degrades substantially across all models (between -0.153 and -0.356 F1). **DistilBERT** suffers the largest drop (F1 $= 0.587$, $\Delta = -0.356$), while **RoBERTa** is most resilient (F1 $= 0.792$, $\Delta = -0.153$). The dominant failure mode is a **recall collapse**: Precision remains stable (>0.90) while Recall plummets (DistilBERT: $0.938 \rightarrow 0.428$), indicating models miss obfuscated hate speech rather than mislabel neutral content — a particularly dangerous failure for deployed moderation systems. This is primarily a *tokenisation problem*: substitution and spacing decompose toxic tokens into rare BPE subword fragments (Sennrich et al., 2016) that carry no toxicity signal, rendering harmful content invisible to the classifier. Elongation — the only technique preserving subword boundaries — is consistently easiest to detect (F1: 0.808–0.878), directly confirming this hypothesis. Comparing D_b and D_c (ΔF1 $\leq$ 0.003), neutral-token abbreviation contributes negligibly to degradation, confirming models respond to *semantic toxicity signals* rather than surface obfuscation patterns — a confound that prior benchmarks (Gröndahl et al., 2018) could not rule out.

5.2 Per-Technique Vulnerability

Table 7 reveals pronounced variation across techniques. **Substitution** is most effective (DistilBERT: F1 $= 0.612$; HateBERT: 0.657), as symbol replacement directly disrupts subword segmentation. **Masking** causes similar degradation for weaker models yet RoBERTa handles it well (F1 $= 0.838$), suggesting its aggressive pre-training (Liu et al., 2019) yields more robust subword representations. **Elongation** is the weakest attack (F1: 0.808–0.878) as character repetition preserves recoverable subword structure. **Composite** chaining produces intermediate degradation, demonstrating that technique combination does not simply accumulate individual effects. Notably, **HateBERT** does not outperform general-purpose encoders despite domain specialisation (ΔF1 $= -0.292$ vs RoBERTa's -0.153), extending Röttger et al. (2021) and Aggarwal and Zesch (2022): domain adaptation improves recognition of *known* toxic surface forms but confers no advantage against deliberate surface disruption. **Robustness requires surface invariance, not domain familiarity.**

5.3 Intensity Analysis

Figure 3 reveals a **non-monotonic** degradation pattern. The sharpest F1 drop occurs at low intensity (1 token), reflecting the *toxic anchor effect*: classifiers rely disproportionately on one or two salient toxic tokens, so corrupting even one collapses detection confidence (DistilBERT: -0.240). At medium intensity (2–3 tokens), stronger models partially recover as structural anomalies — unusual punctuation density, alternating symbol-letter sequences — provide indirect toxicity signals. At high inten-

Dataset	Metric	DistilBERT	RoBERTa	HateBERT	XLM-R
D_a (Clean)	Accuracy	0.994†	0.944	0.945	0.935
	Precision	0.949	0.933	0.943	0.938
	Recall	0.938	0.958	0.947	0.934
	F1	0.943	**0.945**	**0.945**	0.936
D_b (Mixed Obf.)	Accuracy	0.698	0.819	0.731	0.778
	Precision	0.932	0.933	0.921	0.914
	Recall	0.428	0.687	0.506	0.614
	F1	0.587	**0.792**	0.653	0.735
	ΔF1 vs D_a	−0.356	−0.153	−0.292	−0.201
D_c (Realistic Obf.)	Accuracy	0.695	0.818	0.732	0.778
	Precision	0.928	0.934	0.923	0.915
	Recall	0.423	0.684	0.507	0.613
	F1	0.582	**0.790**	0.655	0.734
	ΔF1 vs D_a	−0.361	−0.155	−0.290	−0.202

Table 6: Main results across clean and obfuscated conditions. ΔF1 vs D_a shows drop from clean baseline, shaded light-to-deep red proportional to magnitude. Best F1 per dataset **bolded**. †DistilBERT accuracy reflects a slight positive-class prediction bias; F1 remains the primary metric.

Technique	DistilBERT	RoBERTa	HateBERT	XLM-R
Substitution	0.612	0.770	0.657	0.713
Phonetic	0.766	0.765	0.774	0.779
Spacing	0.686	0.761	0.729	0.757
Masking	0.640	0.838	0.683	0.789
Elongation	0.808	0.878	0.825	0.870
Composite	0.723	0.837	0.756	0.816

Table 7: F1 scores per obfuscation technique (D_d, D1–D6). Cells shaded red-to-teal: **red** = lower F1, **teal** = higher F1.

sity (≥4 tokens), *signal saturation* drives further recovery (RoBERTa: F1 = 0.905), as pervasive surface corruption renders comments structurally anomalous in ways detectable independently of toxic token recognition. Critically, **moderate obfuscation (1–3 tokens) is the most dangerous evasion regime**: a sophisticated adversary maximises evasion by surgically targeting only the most salient toxic tokens, staying below the threshold at which structural anomaly detection activates. This intensity axis was entirely absent from prior benchmarks (Gröndahl et al., 2018; Aggarwal and Zesch, 2022).

5.4 Obfuscation-Aware Training

Table 8 shows that obfuscation-aware fine-tuning ($D_b \rightarrow D_b$) recovers substantial performance across all models, with all recovering to within 0.03 F1 of their clean baseline. DistilBERT achieves the largest gain (+0.326), confirming its vulnerability stems from distributional mismatch rather than architectural limitations. Unlike gradient-based adversarial training (Ebrahimi et al., 2018; Wallace et al., 2019), our approach requires only synthetically generated obfuscated data — producible from any existing hate speech corpus without human annotation. We recommend obfuscation-aware fine-tuning as a **first-line, deployable defence** for any content moderation pipeline, particularly efficiency-constrained systems where DistilBERT's recovery gain is largest.

6 Conclusion

This work presents a structured robustness benchmark for evaluating hate speech detection systems under adversarial text obfuscation. Across six obfuscation techniques, multiple policy regimes,

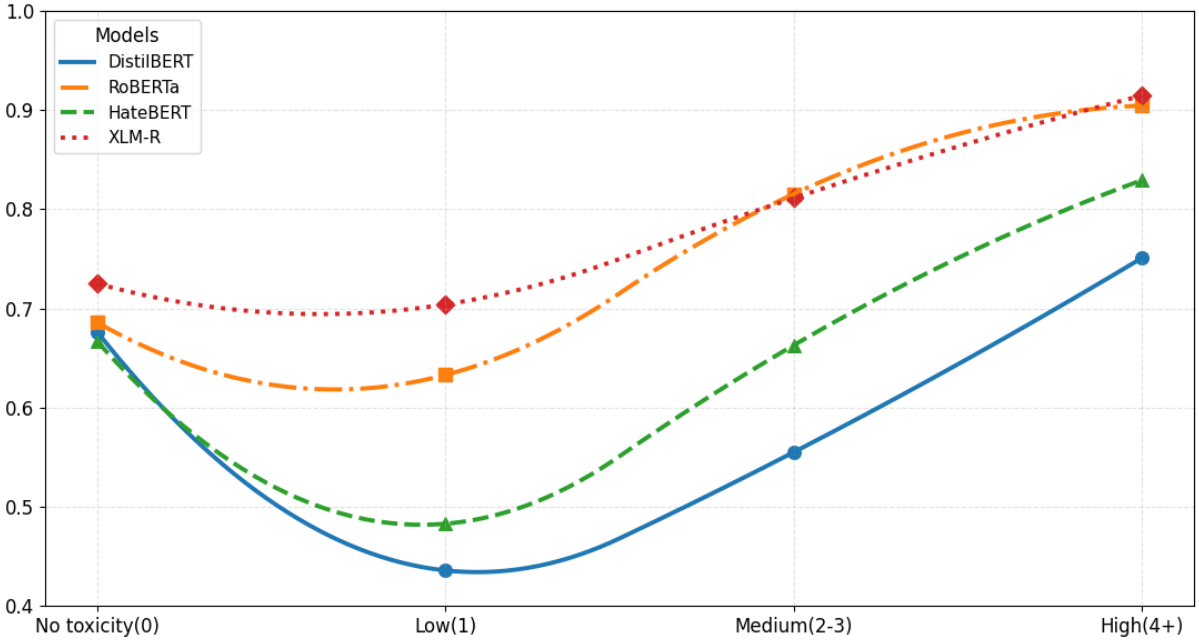

Figure 3: F1 scores across obfuscation intensity bins. Performance is non-monotonic: low intensity (1 token) causes the sharpest drop via the toxic anchor effect, with partial recovery at medium (2–3 tokens) and substantial recovery at high intensity ($\geq$4 tokens) via signal saturation.

Model	$\mathbf{D}_a \rightarrow \mathbf{D}_a$	$\mathbf{D}_a \rightarrow \mathbf{D}_b$	**Δ Degradation**	$\mathbf{D}_b \rightarrow \mathbf{D}_b$	**Δ Recovery**
DistilBERT	0.943	0.587	-0.356	0.913	$+0.326$
RoBERTa	0.945	0.792	-0.153	0.928	$+0.136$
HateBERT	0.945	0.653	-0.292	0.913	$+0.260$
XLM-R	0.936	0.735	-0.201	0.905	$+0.170$

Table 8: F1 performance under clean, obfuscated, and obfuscation-aware training conditions. Δ Degradation and Δ Recovery indicate performance change relative to the clean baseline, with shading proportional to magnitude.

and four transformer architectures, our experiments show that modern hate speech detectors remain highly sensitive to surface-level perturbations despite strong performance on clean benchmarks. Our results demonstrate that even state-of-the-art models are highly vulnerable to deliberate surface-level manipulation, with F1 drops of up to -0.356 driven primarily by tokenisation disruption rather than semantic confusion. Critically, domain specialisation alone does not confer robustness — pre-training breadth and vocabulary coverage are stronger determinants, with RoBERTa consistently outperforming the domain-adapted HateBERT across all conditions. Our intensity analysis identifies moderate obfuscation (1–3 tokens) as the most dangerous evasion regime, a finding absent from prior work. Finally, obfuscation-aware fine-tuning recovers up to $+0.326$ F1 with no architectural changes, offering a practical, scalable, and immediately deployable defence for content moderation systems. We release our dataset variants and obfuscation framework to support reproducible evaluation of future hate speech detection systems under adversarial conditions.

7 Limitations

Synthetic obfuscation. Our obfuscation datasets are algorithmically generated using probabilistic mechanisms rather than collected from real-world adversarial posts. While our policy framework is grounded in documented evasion strategies (Röndahl et al., 2018), synthetically obfuscated text may not fully capture the creativity, context-sensitivity, and linguistic diversity of human-authored evasion.

Obfuscation coverage. Our six techniques, while well-documented, represent a subset of real-world evasion strategies. Emoji-based obfuscation (Kirk et al., 2022), code-switching, homoglyph attacks, zero-width character insertion, and LLM-

generated semantic paraphrases that preserve hateful intent while evading surface-level detection are not covered, and real adversaries may combine these with our studied techniques in unpredictable ways.

Dataset and language scope. Our benchmark derives from a single English Twitter corpus (Davidson et al., 2017) collected in 2017. Hate speech evolves rapidly across platforms and languages, and findings may not generalise to more recent, multilingual, or platform-diverse settings. Obfuscation impact on subword tokenisation may also differ substantially across morphologically rich or non-Latin-script languages.

Model and defence scope. We focus on encoder-only architectures; the robustness of generative LLMs in moderation contexts remains unstudied. Our recovery evaluation also assumes a fixed obfuscation distribution at test time — adaptive adversaries who observe and circumvent the defended model may require more sophisticated defence strategies (Wallace et al., 2019).

8 Ethical Considerations

Dual-use risk. This work studies six textual obfuscation strategies that could potentially be used to evade hate speech detection systems. However, these techniques are already widely documented and commonly observed in real-world online discourse (Gröndahl et al., 2018; Aggarwal and Zesch, 2022). Our contribution is not the creation of new attack mechanisms, but a systematic evaluation framework for measuring model robustness against existing adversarial behaviour. The primary beneficiaries of this work are researchers and platform operators developing more reliable moderation systems.

Dataset release and reproducibility. To support reproducibility and future research in adversarial hate speech detection, we plan to publicly release the obfuscation framework, generated dataset variants, and experimental code upon acceptance of the paper. Although such resources could theoretically be misused to automate toxic-text obfuscation, the techniques studied are already publicly known and widely used online. We believe the scientific value of transparent and standardised robustness evaluation outweighs the associated risks, consistent with open-science principles adopted by the ACL community.

Harmful content. Our benchmark is derived from the dataset of Davidson et al. (2017), which contains real instances of hate speech and offensive language. We do not create new harmful content; all toxic examples originate from the source corpus and are used solely for research and evaluation purposes. Researchers using the dataset should handle it responsibly and follow institutional guidelines for working with sensitive material.

Societal impact. Improving robustness in hate speech detection can help reduce online abuse and better protect marginalised communities. However, automated moderation systems also risk over-censorship, demographic bias, or misuse in restrictive surveillance settings (Vidgen and Derczynski, 2020). We therefore encourage deployment practices that include transparency, human oversight, and regular auditing.

9 Acknowledgements

The authors used AI-assisted tools, including Claude (Anthropic), ChatGPT, and Grammarly, during manuscript preparation for language refinement, literature exploration, and LaTeX formatting assistance. All generated content was carefully reviewed, verified, and revised by the authors prior to submission.

References

Piush Aggarwal and Torsten Zesch. 2022. Analyzing the real vulnerability of hate speech detection systems against targeted intentional noise. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 230–242, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on*

Web and Social Media (ICWSM), volume 11, pages 512–515.

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

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of ACL*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*.

Paula Fortuna and Sérgio Nunes. 2018. A survey of automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. A large scale crowdsourcing dataset for abusive language detection. In *Proceedings of ICWSM*.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of ACL*.

Laura Hanu and Unitary Team. 2020. Detoxify.

Ipsos. 2023. Two in three people often encounter hate speech online.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of ACL*.

Hannah Kirk, Bertie Vidgen, Paul Röttger, and 1 others. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of NAACL-HLT*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Meta Platforms. 2024. Facebook hate speech removal statistics.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2020. Detecting offensive speech in conversational code-switching. In *Proceedings of LREC*.

John Morris, Eli Lifland, Jin Yong Yoo, and 1 others. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of EMNLP: System Demonstrations*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, and 1 others. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of ACL-IJCNLP*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Sali A. Tagliamonte and Derek Denis. 2008. Linguistic ruin? LOL! instant messaging and teen language. *American Speech*, 83(1).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of EMNLP*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. AdvGLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of NeurIPS*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pages 38–45.

A Appendix

A.1 Hyperparameter Settings

All experiments were conducted using pretrained transformer-based models implemented with the HuggingFace Transformers framework and PyTorch on Google Colab (NVIDIA T4 GPU). The models were evaluated using a learning rate of 2×10^{-5}, batch size 16, and maximum sequence length

of 128 tokens. Consistent tokenization, padding, and truncation settings were applied across all experiments to ensure fair comparison between clean and obfuscated hate speech datasets. F1 Score was used as the primary evaluation metric.