Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Exploiting Explainability to Design Adversarial Attacks and Evaluate Attack Resilience in Hate-Speech Detection Models

Pranath Reddy Kumbam[1], Sohaib Uddin Syed[1], Prashanth Thamminedi[1], Suhas Harish[1], Ian Perera[2], Bonnie J. Dorr[1],

[1]University of Florida
[2]Institute for Human and Machine Cognition
kumbam.pranath@gmail.com, sohaibuddinsyed@ufl.edu, pthamminedi@ufl.edu, sganjalaguntehar@ufl.edu
iperera@ihmc.org, bonniejdorr@ufl.edu

Abstract

The advent of social media has given rise to numerous ethical challenges, with hate speech among the most significant concerns. Researchers are attempting to tackle this problem by using hate-speech detection and employing language models to automatically moderate content and promote civil discourse. Unfortunately, recent studies have revealed that hate-speech detection systems can be misled by adversarial attacks, raising concerns about their resilience. While previous research has separately addressed the robustness of these models under adversarial attacks and their explainability, there has been no comprehensive study exploring their intersection. The novelty of our work lies in combining these two critical aspects, leveraging explainability to identify potential vulnerabilities and enabling the design of targeted adversarial attacks. This paper quantifies the interplay between explainability and adversarial robustness in hate-speech detection models. We define novel metrics based on explainability-driven adversarial attacks to evaluate this relationship, providing a clear assessment of model vulnerabilities and guiding the development of more resilient systems.

1 Introduction

Due to the growing influence of social media, understanding online interactions and addressing offensive or hateful content has become increasingly important. Prior work focuses on content moderation (Srinivasan et al. 2019), whereas recent efforts are shifting towards automated mediation to promote civil discourse (Bose, Perera, and Dorr 2023), rather than simply removing offensive posts (Kirk et al. 2022). Hate speech detection is critical for achieving this goal.

Hate speech often touches on sensitive topics such as race, gender, ethnicity, and religion. Models like BERT (Devlin et al. 2018), LSTM (Hochreiter and Schmidhuber 1997), and CNN (Kim 2014) have been used to detect such content, but they struggle with even slightest input perturbations (Goodfellow, Shlens, and Szegedy 2014; Balkir et al. 2022; Ye, Le, and Lee 2023). Notably, such approaches are vulnerable to adversarial attacks that manipulate inputs (e.g., word order changes or synonym replacements) to cause misclassification. These attacks degrade model performance without altering human perception. Automated moderation relies on sophisticated models that analyze and classify content based on intricately predefined rules and guidelines (Ye et al. 2023; Wang et al. 2022).

An important step toward addressing these challenges is understanding the relationship between explainability and prediction accuracy under adversarial conditions. Model explainability refers to the ability to explain and understand the model's decision-making process which helps identify biases, flaws, and vulnerabilities, which is vital for building trust, accountability, and more resilient systems. For instance, consider these two sentences:

- **Sentence A:** "As a woman you shouldn't complain about cleaning up your house, as a man you should always take the trash out..."
- **Sentence B:** "As a woman you shouldn't complain about cleaning up your house, as a man you should permanently take the garbage out..."

Sentence A might be classified as offensive because the model focuses on key terms like 'trash' which heavily influence its decision-making process. However, a slight perturbation in Sentence B, such as substituting 'trash' with a synonym like 'garbage', might lead the model to misclassify it as non-offensive. This confusion arises because the subtle wording change does not alter the intent, yet models often focus on specific keywords or patterns rather than grasping the broader context. This highlights how exposing a model's decision-making process and introducing minor variations can lead to misclassifications or model failures.

Our study focuses on the interplay between explainability and robustness to adversarial attacks. While explainability is essential for understanding model behavior, as seen in the previous example, it can also expose vulnerabilities that adversarial attacks can exploit. We hypothesize that improved explainability may inadvertently expose vulnerabilities, reducing adversarial robustness. By systematically evaluating this trade-off, we aim to introduce a quantitative framework for understanding the behavior of different model architectures and guiding the development of more resilient models. Our key contributions are:

- A novel framework that combines explainability and adversarial robustness in hate-speech detection models, providing a quantitative evaluation of their interplay

- Introduction of explainability-driven adversarial attacks to systematically uncover vulnerabilities in models like BERT, LSTM, and CNN, demonstrating how explainability can be leveraged to design targeted attacks
- Design and validation of novel metrics such as Degree of Explainability (DoE), Adversarial Robustness (AdvRob), and Attack Resilience (Ar), highlighting trade-offs between model explainability and robustness.

While previous studies have explored adversarial robustness (Hauser et al. 2021) and model explainability (Mehta and Passi 2022) independently, our work distinguishes itself by integrating both aspects. We argue that understanding how explainability can be exploited in adversarial attacks is crucial for developing robust hate-speech detection systems that foster safer online environments. In hate-speech detection, explainability helps identify biases and vulnerabilities, guiding the development of more robust and resilient systems. Techniques like LIME (Tulio Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) enhance explainability by explaining predictions, contributing to improved resilience against adversarial attacks.

The next section reviews related work on hate-speech detection and model explainability. Section 3 describes our framework for adversarially robust hate-speech detection. Sections 4 and 5 present our datasets and pre-processing, followed by a description of the models underlying our approach. Explainability for hate-speech detection is introduced in Section 6, followed by a presentation of our explainability-driven adversarial attack algorithm in Section 7. Section 8 provides illustrative examples of these attacks. Section 9 introduces our metrics and evaluation. Following this, we present our results and concluding remarks.

2 Related Work

The rise of online platforms has amplified both freedom of speech and the challenges of racism and hate speech (Williams et al. 2020). Identifying hateful content is crucial for promoting civil discourse, and models like BERT, CNN, and BiRNN have shown success in hate-speech detection (Mathew et al. 2021). However, these models are vulnerable to adversarial attacks and perturbations (Zhang et al. 2020).

Hate-speech detection is vital for content moderation, but even advanced and complex models face significant limitations. Traditional evaluations focus on performance metrics like accuracy, which provide limited insights into model weaknesses (Vidgen and Derczynski 2021). For instance, a model with high accuracy may still exhibit bias against certain groups (Niven and Kao 2019).

Additionally, models may perform well on test sets due to biases in the training data. For example, a model trained on data targeting a specific community may inaccurately flag non-hateful content related to that community as hate speech (Park, Shin, and Fung 2018). Such limitations are problematic given the variability in hate-speech datasets, which differ in sources, sampling strategies, and annotation processes (Kennedy et al. 2020a). These datasets often exhibit annotator and author biases, leading to models that over-rely on lexical features and generalize poorly to other datasets.

Comprehensive evaluation methods that account for these limitations are needed. Beyond performance metrics, enhancing model explainability is crucial for ensuring fairness and transparency in decision-making (Mather et al. 2022). Techniques like LIME (Tulio Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) have been widely used to reveal key features impacting model predictions (Beltagy, Peters, and Cohan 2020).

Adversarial tools like TextAttack (Morris et al. 2020) help assess model vulnerabilities by conducting adversarial attacks, highlighting the need for evaluating models against various perturbations. While Mozafari, Farahbakhsh, and Crespi (2020) analyze BERT for hate-speech detection, there is still a need for evaluating these models across a broader range of perturbations and for exploring the link between explainability and adversarial robustness.

3 Toward Explainable, Adversarially Robust Hate-Speech Detection

We adopt a systematic approach to evaluate the explainability and adversarial robustness of hate-speech detection models, including BERT, LSTM, and CNN. Our study focuses on quantifying the relationship between Explainability and Adversarial Robustness to better understand model vulnerabilities and resilience. To assess explainability, post-hoc techniques (LIME and SHAP) identify key features driving model predictions. This analysis reveals potential biases and vulnerabilities that adversarial attacks could exploit.

Building on these insights, we generate targeted adversarial perturbations, such as synonym substitution, character modifications, and paraphrasing, to test model robustness. We assess model resilience to adversarial attacks by measuring accuracy and other metrics on perturbed datasets.

To further explore the relationship between explainability and robustness, we introduce novel metrics: Degree of Explainability (DoE), Adversarial Robustness (AdvRob), and Attack Resilience (Ar). These metrics help us understand the trade-offs between explainability and robustness, guiding the development of more resilient and explainable hate-speech detection models.

Our approach offers a comprehensive framework for assessing and improving the reliability of hate-speech detection systems. In the following sections, we detail the datasets, pre-processing procedures, model training, and experimental results.

4 Datasets and Pre-processing

We utilize two diverse datasets that capture hate speech prevalence on social media, providing a broad and representative sample of hate speech and offensive language. These datasets, sourced independently and covering varied content, enable a comprehensive evaluation of the models' effectiveness and generalizability across different contexts.

Hate Speech and Offensive Language

Kaggle's Hate Speech and Offensive Language dataset (Samoshyn 2020) focuses on hate speech and offensive language usage on Twitter.

This dataset possesses several key attributes: (a) it includes the tweet text, representing the original, unprocessed textual data from Twitter; (b) Each tweet in the dataset is assigned one of three class labels: hate speech, offensive language, or neither. The labels originate from human expert annotations made during the dataset's initial collection. (c) its 24,783 unique entries make it an abundant resource for training and evaluating our models. We leverage the multi-class structure to assess models' performance across a spectrum of offensive content, spanning from blatant hate speech to subtler forms of offensive language. The dataset is abbreviated as HSOL (Hate Speech and Offensive Language).

UC Berkeley Measuring Hate-speech The publicly released Berkeley Measuring Hate-speech dataset, previously used in an experiment on hate-speech detection (Kennedy et al. 2020b), has several key features that complement those of the first dataset: (a) it comprises over 135,556 combined rows of data, with 39,565 unique comments that have been annotated with ordinal labels by 7,912 annotators; (b) it is designed for binary classification, with only two categories: "Hate speech" or "Not hate speech;" (c) the dataset's main outcome variable is the hate-speech score, ranging from 0 to 1, where $>$0.5 denotes hate speech, and $\leq$0.5 denotes *not* hate speech. The binary classification enables a focused evaluation of models' performance in accurately detecting and classifying hate speech instances, simplifying our assessment. Each is described, in turn, below. This dataset will be abbreviated as BHS (Berkeley Hate Speech). The BHS dataset comprises social media posts collected from multiple platforms, including YouTube, Reddit, and Twitter. Each instance in the dataset represents an individual post, annotated with its corresponding label.

Pre-processing

To ensure the quality and consistency of the data, we perform several pre-processing steps on both datasets. These steps include converting the tweets to lowercase, removing punctuation, extra spaces, URLs, mentions, and hashtags. Tokenization is performed using the NLTK library (Bird and Loper 2004) for the CNN and LSTM models. For the implementation of DistilBERT, we use the DistilBertTokenizer from the HuggingFace transformers library to tokenize the text data.

Following tokenization, we perform lemmatization to convert the words to their base forms, which aids in reducing dimensionality and improving the generalizability of the models. To further reduce noise and focus on meaningful words, we remove stop words in tweets.

These pre-processing steps ensure that the input data for the hate-speech detection models are clean, consistent, and representative of the underlying language patterns and structures.

5 Models

We select DistilBERT (Sanh et al. 2019), LSTM (Hochreiter and Schmidhuber 1997), and CNN (Bengio and Lecun 1997) to explore three distinct paradigms in machine learning for handling text data. DistilBERT, a distilled version of BERT (Devlin et al. 2018), benefits from pre-training on large corpora, allowing it to capture complex linguistic patterns effectively. LSTM specializes in sequential data, making it well-suited for capturing temporal dependencies, while CNN excels in local feature detection, focusing on extracting relevant features from fixed-size windows.

We employ the AdamW optimizer (Loshchilov and Hutter 2017) for model training, with a learning rate of $2e^{-5}$. To further optimize the training process, we apply a linear learning rate scheduler with warmup. The training involves iterative updates to the model's parameters using mini-batch gradient descent over a fixed number of epochs, with performance monitored using the average training loss per epoch. The models are implemented using the `PyTorch` package (Paszke et al. 2019) and trained on an NVIDIA Tesla T4 GPU.

We selected LSTM, CNN, and Distilled BERT as representative models for this study. These architectures span a range of design paradigms commonly used in hate-speech detection. LSTM captures sequential dependencies, CNN excels at extracting local features, and Distilled BERT represents modern transformer-based pre-trained models with computational efficiency. Together, these models enable a balanced evaluation of the interplay between explainability and robustness across diverse architectures.

6 Determining Explainability for Hate-Speech Detection

Our experimentation is designed to assess the explainability of the hate-speech detection models we have trained (LSTM, CNN, and DistilBERT). The goal is to gain a deeper understanding of how these models make decisions and provide insights into their decision-making processes. Through this evaluation, we expect to gain valuable insights into the inner workings of these models and identify potential areas for improvement. We employ the LIME technique, which aids in interpreting and explaining the models' predictions.

Explainability

Our research employs both LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to provide explanations for the predictions made by our hate-speech detection models, independent of their underlying architecture. LIME generates localized explanations by approximating the model's behavior near a specific instance using a simpler, explainable model, such as a linear model. Similarly, SHAP assigns consistent importance scores to each feature, offering a unified measure of feature importance across different models.

To apply LIME, we first select an instance and perturb it, then obtain model predictions for these perturbed instances. Weights are assigned based on similarity to the original instance, and a simpler model is trained to approximate the model's behavior around that instance. The output is a set of token-level explanations, with each token's contribution quantified by a score $(T, < score >)$.

Algorithm 1: Explainability-Driven Adversarial Attack

Require: Input sentence S, model M, feature importance scores F, synonym dictionary D

$T \leftarrow$ Rank features based on scores in F
for each feature $t \in T[1:K]$ **do**
  if t is a word and $D[t]$ exists **then**
    Replace t in S with its synonym $D[t]$
  else
    Perform character-level modification on t (replace, add, or delete a character)
  end if
  $S' \leftarrow$ modified sentence
  $y' \leftarrow M(S')$ {Evaluate modified sentence with model}
  if $y' \neq y$ **then**
    Stop {Successful attack if classification changes}
  end if
end for
return S', y'

By using LIME and SHAP, we gain valuable insights into the features that influence the model's predictions, helping us evaluate their decision-making processes. This understanding is crucial for building trust in the models, particularly in sensitive domains like hate-speech detection, where transparency and reliability are essential.

7 Adversarial Attacks

After assessing the explainability of each hate-speech detection model (LSTM, CNN, and DistilBERT) using LIME and SHAP, we introduce adversarial attacks designed to modify the original text while preserving its intended meaning. Our approach involves a greedy method that ranks features based on their importance scores from LIME or SHAP. We then perform targeted adversarial attacks on the K most significant features.

For each selected feature, if it is a word with a suitable synonym, we replace it with that synonym. If the feature is a token or lacks an appropriate synonym, we perform character-level modifications such as replacement, addition, or deletion. The modified sentence is then re-evaluated by the model to assess its robustness. This method allows us to systematically test the model's resilience against perturbations targeting its most influential features.

The effectiveness of these adversarial attacks is assessed by comparing the model's classification results before and after the modification. By analyzing the variations in classification, we gain insights into the vulnerabilities of the models and their ability to withstand adversarial manipulations. The detailed steps of this process are outlined in Algorithm 1. For this study, we set K to 2, meaning the two most significant features are modified for adversarial attacks.

Through the use of explainability-driven adversarial attacks, we assess the robustness of our hate-speech detection models against malicious manipulations. By identifying weaknesses and improving the models, we can enhance their resilience against adversarial attacks and ensure their reliability in real-world applications.

Our choice of synonym and character-level substitutions as adversarial attack strategies is intentional and closely aligned with the goals of this study. These attacks exploit vulnerabilities revealed through explainability techniques like LIME and SHAP, enabling us to study the unique interplay between explainability and adversarial robustness. While other types of attacks, such as sentence paraphrasing (Iyyer et al. 2018) or grammar-based perturbations, (Dong et al. 2023) are commonly employed in NLP research, they do not directly leverage explainability insights and are therefore outside the scope of this work. By focusing on explainability-driven attacks, we aim to provide a deeper understanding of how explainability can expose model vulnerabilities.

While our attack algorithm uses basic synonym and character replacements, it could be enhanced with more sophisticated, context-aware modifications targeting words around the most significant or harmful features (see Section 8). This is a limitation of our study (see Section 12), to be addressed by exploring more advanced attacks in future work.

8 Adversarial Attack Examples

Example 1 illustrates a successful attack where a minor modification to the sentence leads the LSTM model to misclassify the input as neither offensive nor hate-speech, despite the appearance of a derogatory term (b***h). This example illustrates a deviation in LIME results between the original and modified sentences, highlighting the increased susceptibility of the sentence to adversarial attacks. The low scores assigned by LIME to the words in the perturbed sentence indicate that none of these words significantly contribute to classifying the sentence as hate speech, reflecting a successful adversarial attack.

Example 1: Offensive Language Retained

Ground Truth Label: 1

Before Attack:

Original Sample: @rhythmixx_ :hobbies include: fighting Mariam b***h

Original LIME Scores: hobbies (0.03134), include (0.02144), fighting (0.05365), Mariam (0.04012)

After Attack:

Perturbed Sample: @rhythmixx_ :hobbies including: struggle Mariam b***h

Perturbed LIME Scores: hobbies (0.01293), including (0.01102), struggle (-0.01891), Mariam (-0.00517)

Attack Result: 1 (90%) $\rightarrow$ 0 (34%)

Example 2: Gender Stereotype Retained

Ground Truth Label: 1

Before Attack:

Original Sample: As a woman you shouldn't complain about cleaning up your house, as a man you should always take the trash out...

Original LIME Scores: "trash" (0.09836), "always" (-0.06172), "you" (0.04416), "complain" (0.03224), "should" (0.03076)

After Attack:

Perturbed Sample: As a woman you shouldn't complain about cleaning up your house, as a man you should permanently take the garbage out...

Perturbed LIME Scores: "permanently" (0.02808), "man" (-0.02807), "you" (0.01809), "garbage" (0.01683), "complain" (-0.01672)

Attack Result: 1 (92%) $\rightarrow$ 0 (51%)

Example 2 presents another case where a slight change in the sentence causes the model to classify it as non-offensive. This sentence promotes damaging gender stereotypes, suggesting that it is offensive and warrants a label of 1, not 0.

The examples in this section are provided for illustrative purposes to help readers understand the general principles behind adversarial attacks. These examples do not necessarily follow the proposed attack algorithm described in Algorithm 1, which selects features strictly based on their LIME/SHAP scores and their contribution to the model's prediction. LIME and SHAP scores quantify the contribution of individual features to the model's prediction. Positive scores indicate features that support the predicted class, while negative scores indicate features that detract from it.

9 Metrics and Evaluation

Evaluating the performance of hate-speech detection models is crucial to ensuring their reliability and effectiveness in real-world applications. By obtaining the explainability results from LIME and the respective attack outcomes for various models, we establish a comprehensive framework to assess their Adversarial Robustness. We propose the metrics below to gauge their performance.

Degree of Explainability

LIME offers a valuable tool for assessing the importance of individual words in a sentence when it comes to the classification decisions made by the model. By providing a dictionary of values, known as prediction probabilities, we assess the impact of each word in determining whether a sentence qualifies as hate speech or not.

To quantify the Degree of Explainability, we first calculate the standard deviation of the prediction probabilities for each sentence. Next, we determine the fraction of words whose prediction probability surpasses the standard deviation.

$$\text{DoE} = \frac{\text{number of features with score} > \sigma}{\text{total number of features}} \quad (1)$$

A high Degree of Explainability indicates that the model's classification decision is primarily driven by a few words with substantial explainability scores. As a result, altering these words could substantially change the explainability outcome, rendering the sentence more vulnerable to adversarial attacks.

Adversarial Robustness

To measure a model's resilience against adversarial attacks, we quantify the Adversarial Robustness by computing the ratio between the model's accuracy under attack and its accuracy before the attack. A ratio near 1 signifies a high level of robustness, indicating that the model can withstand adversarial manipulations. Conversely, a lower ratio suggests an increased vulnerability to adversarial attacks.

$$\text{AdvRob} = \frac{\text{accuracy after attack}}{\text{accuracy before attack}} \quad (2)$$

Attack Resilience

Understanding the relationship between Adversarial Robustness and Degree of Explainability is crucial for developing more effective and reliable hate-speech detection models. We compute the ratio of Adversarial Robustness to Degree of Explainability, which we define as Attack Resilience, to assess this relationship.

$$\text{Ar} = \frac{\text{AdvRob}}{\text{DoE}} \quad (3)$$

A high Degree of Explainability suggests that a model is easily interpretable, but it also makes it more susceptible to adversarial attacks, resulting in lower Adversarial Robustness and Attack Resilience. Conversely, a low Degree of Explainability means that the model is less interpretable but more resistant to adversarial attacks, yielding higher Attack Resilience. The ideal approach is to find a balance between explainability and resilience to achieve the best Attack Resilience.

By using these metrics, we can evaluate the performance of various hate-speech detection models, identify areas for improvement, and guide the development of more robust models for real-world applications. This comprehensive analysis empowers researchers and practitioners to make informed decisions regarding the design and deployment of hate-speech detection models. Consequently, it supports the development of more effective solutions for addressing online hate speech while ensuring resilience against adversarial perturbations.

10 Results

We evaluate the performance of each language model—DistilBERT, CNN, LSTM—using Accuracy (A), Precision (P), Recall (R), F1 Score, and Area Under the Receiver Operating Characteristic Curve (AUC) on two datasets.

These metrics provide a comprehensive assessment of each model's ability to detect hate speech and offensive language. The results are then used to compare the performance of each model against the others.

Table 1 presents our experiment results for multi-class classification on Kaggle's multi-class "Hate Speech and Offensive Language Datset." Table 2 presents the corresponding results for binary classification, on the "Measuring Hate-Speech Dataset."

Model	A	P	R	F1	AUC
DistilBERT	0.91	0.91	0.91	0.91	0.94
LSTM	0.88	0.86	0.88	0.87	0.91
CNN	0.87	0.85	0.87	0.86	0.88

Table 1: Performance Comparison on Kaggle's "Hate Speech and Offensive Language Dataset."

Model	A	P	R	F1	AUC
DistilBERT	0.97	0.97	0.97	0.97	0.99
LSTM	0.96	0.96	0.96	0.96	0.99
CNN	0.96	0.96	0.96	0.96	0.98

Table 2: Performance Comparison on UC Berkeley's "Measuring Hate-Speech Dataset."

While DistilBERT consistently outperforms across all metrics, as expected given its greater model complexity, it is noteworthy that both CNN and LSTM models deliver competitive performance. Their scores are close to those of DistilBERT, making our study of the tradeoff between explainability and robustness more balanced and fair.

Explainability

In this section, we assess the explainability of our hate-speech detection models—LSTM, CNN, and DistilBERT—using both LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). The goal is to quantify the Degree of Explainability (DoE) for each model, which provides insight into how interpretable the the models' decision-making processes are and identifies the most influential in their predictions.

Figures 1 and 2 show the Degree of Explainability (DoE) for different models, as measured by LIME and SHAP, alongside their complexity ($\#Params$) represented in corresponding histograms. Figure 1 illustrates that the LSTM model has a higher DoE according to LIME, indicating that its predictions are more heavily influenced by specific tokens or features. Similarly, Figure 2 shows that LSTM maintains a higher DoE when assessed with SHAP, indicating consistent findings across different explainability methods.

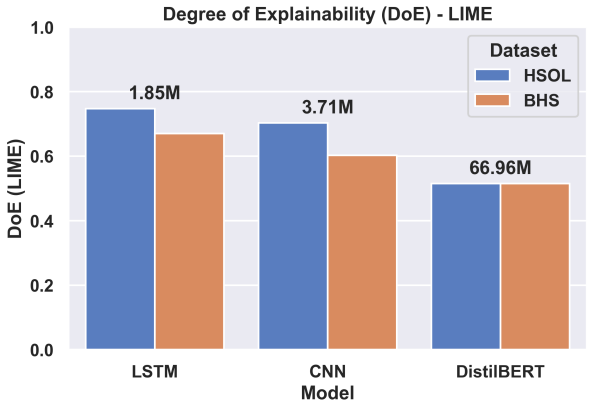

Figure 1: Degree of Explainability (LIME)

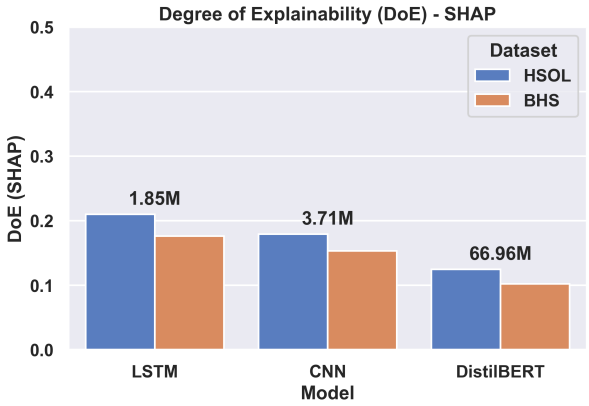

Figure 2: Degree of Explainability (SHAP)

Model	Dataset	DoE	Params (M)
LSTM	HSOL	0.74708	1.85
CNN	HSOL	0.70319	3.71
DistilBERT	HSOL	0.51551	66.96
LSTM	BHS	0.67	1.85
CNN	BHS	0.60253	3.71
DistilBERT	BHS	0.515	66.96

Table 3: Degree of Explainability (DoE) measured using LIME for different models across HSOL and BHS datasets.

Tables 3 and 4 present the quantitative DoE values for different models across the two datasets (HSOL and BHS). From Table 3, we observe that LSTM exhibits a high DoE, while DistilBERT, despite its complex architecture, has the lowest DoE among the models. This suggests that DistilBERT's decisions are less reliant on specific tokens, potentially making it less explainable but more robust.

Additionally, we notice an inverse correlation between model complexity (measured by the number of parameters) and DoE. DistilBERT, with the most parameters, shows the lowest DoE, while LSTM, with the fewest parameters, has

Model	Dataset	DoE	Params (M)
LSTM	HSOL	0.21006	1.85
CNN	HSOL	0.17924	3.71
DistilBERT	HSOL	0.125	66.96
LSTM	BHS	0.1765	1.85
CNN	BHS	0.15329	3.71
DistilBERT	BHS	0.10236	66.96

Table 4: Degree of Explainability (DoE) measured using SHAP for different models across HSOL and BHS datasets.

the highest. This suggests that more complex models may offer greater predictive power at the expense of explainability, which is a crucial consideration when deploying these models in sensitive applications like hate-speech detection.

Category	Percentage	L	C	B
Total	100%	0.67	0.6025	0.515
Disability	2.65%	0.4983	0.4399	0.3795
Religion	19.07%	0.645	0.5639	0.4889
Gender	29.94%	0.6585	0.5527	0.4893
Race	35.69%	0.6533	0.6096	0.5116

Table 5: Degree of Explainability (LIME) by Category

Table 5 provides the Degree of Explainability (DoE) by category (e.g., race, gender, religion) for the LSTM (L), CNN (C), and DistilBERT (B) models and the composition of samples in the dataset corresponding to each of the target category. Notably, categories representing minority target groups, such as "Disability," make up only a small percentage of the total samples in the dataset. This lower representation correlates with a reduced DoE in these categories, which suggests that the models are less explainable when dealing with these minority groups.

This reduced explainability is concerning, as it suggests models may struggle to provide transparent and reliable decisions for these groups. The lack of explainability could lead to biased outcomes, highlighting the need for targeted efforts to enhance the DoE in underrepresented categories. Ensuring that models remain explainable across all demographic groups is crucial for creating fair and trustworthy hate-speech detection systems.

These findings highlight the trade-offs between model complexity and explainability. While simpler models like CNN and LSTM provide more explainable decisions, they might also be more vulnerable to adversarial attacks, as we will see in the subsequent section.

Adversarial Attacks

After determining the Degree of Explainability for each model, we introduce targeted adversarial attacks to evaluate their robustness. Our approach leverages the insights gained from LIME to identify the most influential features in a model's decision-making process. We then generate adversarial perturbations by modifying these critical features to test the model's resilience.

Our experiments evaluate adversarial attacks that alter model predictions in either direction, e.g., from 'toxic' to 'not toxic' or vice versa. Both directions represent valid adversarial scenarios: the former highlights a model's vulnerability to underestimating harmful content, while the latter illustrates over-sensitivity, potentially leading to false positives. By analyzing both directions, our metrics provide a holistic assessment of model performance under adversarial conditions.

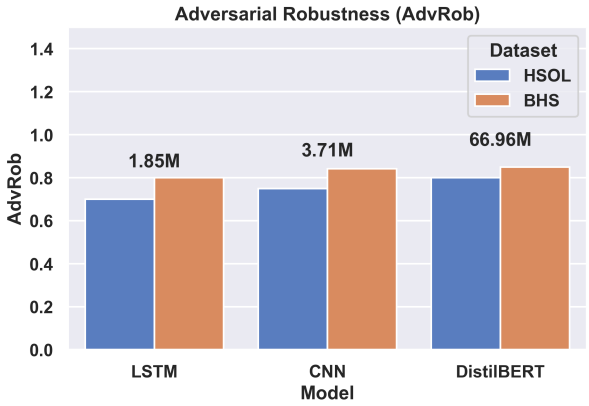

Figure 3: Adversarial Robustness (AdvRob) of different model

Figure 3 shows the Adversarial Robustness (AdvRob) of different models across the two datasets. The results confirm that models with a lower DoE, such as DistilBERT, tend to exhibit higher robustness against adversarial attacks. DistilBERT's robustness can be attributed to its less explainable decision-making process, which makes it harder for adversaries to identify and manipulate key features.

Model	Dataset	AdvRob	Params (M)
LSTM	HSOL	0.7	1.85
CNN	HSOL	0.75	3.71
DistilBERT	HSOL	0.8	66.96
LSTM	BHS	0.8	1.85
CNN	BHS	0.8421	3.71
DistilBERT	BHS	0.85	66.96

Table 6: Adversarial Robustness (AdvRob) of different models across HSOL and BHS datasets.

Table 6 presents the quantitative measures of Adversarial Robustness. We see that DistilBERT maintains a high level of accuracy under adversarial conditions, while LSTM, with its higher DoE, shows the lowest robustness. This aligns with the notion that higher explainability can lead to greater vulnerability, as adversaries can exploit the features that are crucial for the model's decisions.

Finally, Figure 4 and Table 7 illustrate the Attack Re-

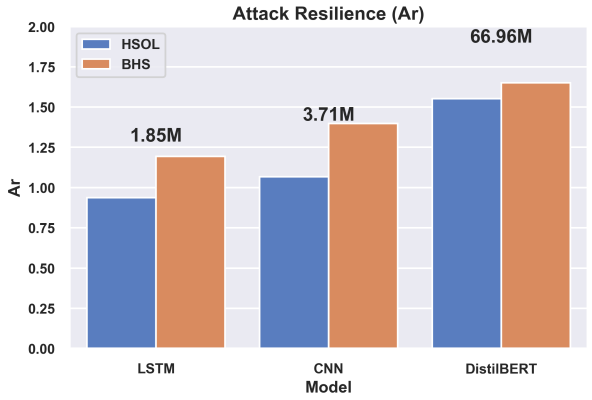

Figure 4: Attack Resilience (Ar) of different models

Model	Dataset	Ar	Params (M)
LSTM	HSOL	0.93697	1.85
CNN	HSOL	1.06656	3.71
DistilBERT	HSOL	1.55186	66.96
LSTM	BHS	1.19402	1.85
CNN	BHS	1.3976	3.71
DistilBERT	BHS	1.65048	66.96

Table 7: Attack Resilience (Ar) of different models across HSOL and BHS datasets.

silience (Ar), which is the ratio of Adversarial Robustness to Degree of Explainability. This metric quantifies the trade-off between explainability and robustness. As shown in Figure 4, DistilBERT has the highest Attack Resilience, indicating a better balance between robustness against adversarial attacks and maintaining some level of explainability. In contrast, LSTM, with its high DoE and lower robustness, has the lowest Attack Resilience, making it more susceptible to adversarial manipulation.

LSTMs can capture some sequential information and dependencies, but their memory capacity is relatively limited compared to the attention mechanism used by BERT. As a result, LSTMs exhibit lower robustness than BERT. CNNs, designed primarily for handling grid-like data such as images, have a more limited capacity to capture contextual information and dependencies in sequential data like text. Their focus on local features makes them vulnerable to adversarial attacks.

These results underscore the importance of carefully balancing explainability and robustness when developing hate-speech detection models. By understanding these trade-offs, we can design models that are not only explainable but also resilient to adversarial attacks, ensuring their effectiveness in real-world applications.

Adversarial Robustness: Distilled BERT vs. BERT Base

To explore the relationship between model complexity and adversarial robustness, we compared two models: Distilled BERT (66M parameters) and BERT Base (110M parameters). Both models were trained under identical configurations: a batch size of 32, 5 epochs, and a learning rate of 2×10^{-5}. Both achieved a similar accuracy of approximately 0.91 on the validation set.

The results, shown in Table **??**, reveal that Distilled BERT exhibits slightly higher adversarial robustness compared to BERT Base across different numbers of modified features (K). For instance, at $K = 2$, the adversarial robustness of Distilled BERT is 0.7826, while that of BERT Base is 0.7733. This trend persists as K increases, highlighting that the smaller Distilled BERT model does not exhibit inferior robustness despite its lower parameter count.

As expected, increasing K leads to a gradual decline in adversarial robustness for both models. For example, robustness for Distilled BERT decreases from 0.7826 at $K = 2$ to 0.7487 at $K = 5$, and a similar trend is observed for BERT Base, which drops from 0.7733 to 0.7518. This demonstrates that explainability-driven adversarial attacks effectively exploit model vulnerabilities irrespective of model size.

The slightly better performance of Distilled BERT, despite its smaller size, could be attributed to the knowledge distillation process, which transfers knowledge from a larger teacher model. This process may enable the distilled model to generalize better and retain critical decision-making patterns, making it more resilient to adversarial perturbations. However, this is only a hypothesis and requires further investigation to confirm. Additionally, the metrics proposed in this work, such as adversarial robustness, enable a quantitative comparison of models, as illustrated here. By systematically evaluating robustness under adversarial attacks, these metrics provide valuable insights into model behavior and facilitate informed decisions about model selection and deployment.

Model	(K)	Adversarial Robustness
BERT Base	2	0.7733
	3	0.7682
	4	0.7549
	5	0.7518
Distilled BERT	2	0.7826
	3	0.7733
	4	0.7672
	5	0.7487

Table 8: Adversarial robustness comparison of BERT Base and Distilled BERT for varying K. Results are averaged over 5 runs on a larger set of 200 samples.

Discussion on Explainability-Robustness Trade-off

An important observation from this study is the inherent trade-off between model explainability and robustness to adversarial attacks. Models with a higher degree of explainability (DoE), such as LSTM, tend to rely more heavily on

specific features for their predictions. While this makes their decision-making processes easier to interpret, it also exposes critical vulnerabilities that adversaries can exploit by targeting these features. In contrast, more complex models, such as Distilled BERT and BERT Base, exhibit lower DoE and higher adversarial robustness (AdvRob), making them less susceptible to such attacks.

This trade-off is quantitatively captured by the Attack Resilience (Ar) metric introduced in this work. As shown in our experiments, models with lower DoE generally achieve higher Ar, indicating that reduced interpretability can contribute to improved robustness. However, this comes at the cost of reduced transparency, which is crucial for trust and accountability in sensitive applications like hate-speech detection.

Balancing this trade-off requires strategic interventions, such as employing ensemble models that combine the strengths of interpretable and robust architectures or utilizing adversarial training to reinforce resilience while maintaining some level of explainability. By systematically analyzing these trade-offs using our proposed metrics, this study provides a framework for developing models that optimize both interpretability and operational resilience.

11 Conclusion and Future Work

This paper investigates the interplay between explainability and adversarial robustness in hate-speech detection models. We conduct attacks on commonly used models like BERT, LSTM, and CNN, focusing on exploiting explainable features to uncover vulnerabilities. Our findings support the hypothesis that enhancing model explainability can increase vulnerability to adversarial attacks, compromising robustness.

We demonstrate a proportional relationship between several factors, including explainability, adversarial robustness, and model complexity, offering insights for fine-tuning models to balance these aspects effectively. By leveraging techniques like LIME and SHAP, we analyze model decision-making and target highly explainable features to generate adversarial examples. Our results show that models with higher explainability are more susceptible to attacks, as adversaries can exploit the key features driving the model's decisions.

This research highlights the importance of balancing explainability and robustness in the design of hate-speech detection models. Achieving this balance ensures that models are not only accurate and explainable but also resilient to adversarial attacks. This approach is essential for developing reliable and effective models, enhancing downstream applications like automated moderation and promoting civil discourse. Employing a combination of techniques, such as adversarial training and ensemble approaches, is likely to improve robustness while maintaining explainability. By understanding and quantifying the trade-offs between these factors, content moderation systems can be designed to balance interpretability with operational resilience effectively.

The current study is limited to models like DistilBERT, CNN, and LSTM, which do not fully capture the capabilities of more advanced Large Language Models (LLMs) and Foundation Models. Future work will extend this analysis to include open source models such as LLaMA (Touvron et al. 2023) and Mistral (Jiang et al. 2023), which offer greater potential for accuracy and robustness in hate-speech detection.

12 Limitations

One limitation of our study is the relatively narrow scope of the models explored. We have focused primarily on CNN, LSTM, and DistilBERT architectures, potentially restricting the generalizability of our findings. More advanced transformer-based architectures such as Large Language Models (LLMs) might exhibit different behaviours in terms of explainability and adversarial robustness.

Our study also focuses on a limited range of adversarial attacks, primarily synonym replacement and character-level modifications, which do not cover more advanced context-aware (Li et al. 2021) or paraphrasing techniques (Iyyer et al. 2018).

By incorporating LLMs, we aim to explore how these models balance explainability and robustness, particularly under more sophisticated adversarial attacks tailored to their unique architectures. This extension will also involve evaluating model performance across a broader range of datasets to ensure generalizability and improve the resilience of hate-speech detection systems.

Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112290022. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Balkir, E.; Nejadgholi, I.; Fraser, K. C.; and Kiritchenko, S. 2022. Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection. *arXiv preprint arXiv:2205.03302*.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv e-prints*, arXiv:2004.05150.

Bengio, Y.; and Lecun, Y. 1997. Convolutional Networks for Images, Speech, and Time-Series. *The handbook of brain theory and neural networks*.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Bose, R.; Perera, I.; and Dorr, B. 2023. Detoxifying Online Discourse: A Guided Response Generation Approach for Reducing Toxicity in User-Generated Text. In Chawla, K.; and Shi, W., eds., *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, 9–14. Toronto, Canada: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, arXiv:1810.04805.

Dong, H.; Dong, J.; Yuan, S.; and Guan, Z. 2023. Adversarial Attack and Defense on Natural Language Processing in Deep Learning: A Survey and Perspective. In Xu, Y.; Yan, H.; Teng, H.; Cai, J.; and Li, J., eds., *Machine Learning for Cyber Security*, 409–424. Cham: Springer Nature Switzerland.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, arXiv:1412.6572.

Hauser, J.; Meng, Z.; Pascual, D.; and Wattenhofer, R. 2021. BERT is Robust! A Case Against Synonym-Based Adversarial Examples in Text Classification. *arXiv e-prints*, arXiv:2109.07403.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1875–1885. New Orleans, Louisiana: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Singh Chaplot, D.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Renard Lavaud, L.; Lachaux, M.-A.; Stock, P.; Le Scao, T.; Lavril, T.; Wang, T.; Lacroix, T.; and El Sayed, W. 2023. Mistral 7B. *arXiv e-prints*, arXiv:2310.06825.

Kennedy, B.; Jin, X.; Mostafazadeh Davani, A.; Dehghani, M.; and Ren, X. 2020a. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5435–5442. Online: Association for Computational Linguistics.

Kennedy, C. J.; Bacon, G.; Sahn, A.; and von Vacano, C. 2020b. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. arXiv:2009.10277.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. *arXiv e-prints*, arXiv:1408.5882.

Kirk, H.; Birhane, A.; Vidgen, B.; and Derczynski, L. 2022. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 497–510. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Li, D.; Zhang, Y.; Peng, H.; Chen, L.; Brockett, C.; Sun, M.-T.; and Dolan, B. 2021. Contextualized Perturbation for Textual Adversarial Attack. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5053–5069. Online: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. *arXiv e-prints*, arXiv:1711.05101.

Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Artificial Intelligence (cs.AI), FOS: Computer and information sciences, Machine Learning (cs.LG), Machine Learning (stat.ML)*. Publisher: arXiv Version Number: 2.

Mather, B.; Perera, I.; Kazakova, V. A.; Capecci, D.; Garg, M.; Woodard, D. L.; and Dorr, B. J. 2022. Vision: Explainable Hidden Mental States as Influence Indicators. In de Melo, P. O. S. V.; Jeng, W.; and Buntain, C., eds., *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media, ICWSM 2022 Workshops, Atlanta, Georgia, USA [hybrid], June 6, 2022*.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14867–14875.

Mehta, H.; and Passi, K. 2022. Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). *Algorithms*, 15(8).

Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, 928–940. Springer.

Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664. Florence, Italy: Association for Computational Linguistics.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799–2804. Brussels, Belgium: Association for Computational Linguistics.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style,

High-Performance Deep Learning Library. *arXiv e-prints*, arXiv:1912.01703.

Samoshyn, A. 2020. Hate Speech and Offensive Language Dataset. *Kaggle*. https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv e-prints*, arXiv:1910.01108.

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. In *Proceedings of the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, arXiv:2302.13971.

Tulio Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*, arXiv:1602.04938.

Vidgen, B.; and Derczynski, L. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12): 1–32.

Wang, H. M.; Bulat, B.; Fujimoto, S.; and Frey, S. 2022. Governing for Free: Rule Process Effects on Reddit Moderator Motivations. In *International Conference on Human-Computer Interaction*, 97–105. Springer.

Williams, M. L.; Burnap, P.; Javed, A.; Liu, H.; and Ozalp, S. 2020. Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1): 93–117.

Ye, M.; Sikka, K.; Atwell, K.; Hassan, S.; Divakaran, A.; and Alikhani, M. 2023. Exploring the Intersection of Explainability and Adversarial Robustness in Hate-Speech Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.

Ye, Y.; Le, T.; and Lee, D. 2023. NoisyHate: Benchmarking Content Moderation Machine Learning Models with Human-Written Perturbations Online. In *Preprint*, 8. New York, NY, USA: ACM.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3): 1–41.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, The research advances scientific understanding of model vulnerabilities without violating social norms or ethics.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, The claims accurately reflect the contributions and scope of the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, The methodological approach aligns with the claims made in the study.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, Artifacts and population-specific distributions are not explicitly clarified.
   (e) Did you describe the limitations of your work? Yes, Please refer to Section 12
   (f) Did you discuss any potential negative societal impacts of your work? Yes, Please refer to the below subsection.
   (g) Did you discuss any potential misuse of your work? Yes, Misuse, especially through adversarial attacks, is acknowledged.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, Please refer to the below subsection.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, The ethics review guidelines were followed.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? No explicit hypothesis testing is involved.
   (b) Have you provided justifications for all theoretical results? No theoretical results requiring justifications are presented.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Competing theories are not part of this paper's focus.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No alternative mechanisms or explanations were necessary.
   (e) Did you address potential biases or limitations in your theoretical framework? Theoretical frameworks are not part of this paper's focus.
   (f) Have you related your theoretical results to the existing literature in social science? This paper is not related to social science literature.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Implications for policy or further social science research are not addressed
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? No theoretical proofs are included.
   (b) Did you include complete proofs of all theoretical results? No proofs are included.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, Instructions are included, but the explicit link to the code is omitted to preserve author anonymity.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, Please refer to Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, Error bars were not included due to page limit constraints.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, Please refer to Section 5
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, The evaluation is justified based on the claims.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, The cost of misclassification is implied in terms of societal impact.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, Please refer to Section 4.
   (b) Did you mention the license of the assets? Licenses are not explicitly mentioned.
   (c) Did you include any new assets in the supplemental material or as a URL? No new assets are released.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, The paper uses publicly available datasets, explicit consent was not required.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, Offensive content is mentioned, but no personally identifiable information is used.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No new datasets are released.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Datasheets are not relevant.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? No crowdsourcing or human subjects are involved.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No IRB approvals were needed.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No compensation was provided.

(d) Did you discuss how data is stored, shared, and deidentified? Data sharing and de-identification are not applicable.

Broader Impact and Ethical Considerations

Our study explores the tradeoff between explainability and adversarial robustness in hate-speech detection models. This research has the potential to deepen our comprehension of the functioning of these models and offer valuable insights for enhancing their design. Additionally, it can contribute to the improvement of downstream tasks like moderation and mediation for promoting ethical discourse, leading to more effective and secure utilization of these models.

The positive impacts of this study include the potential to enhance the performance and security of hate speech detection models, which could lead to more effective and accurate moderation on digital platforms. However, we fully acknowledge the potential for negative outcomes. The adversarial attacks we employ for testing model robustness could, in theory, be misused to weaken the effectiveness of hate speech detection models and propagate harmful content.

It is also important to note that while we strive for model explainability, there is a risk of over-simplifying complex models, which could lead to misinterpretation where the simplified explanation does not accurately reflect the model's behavior.

To mitigate the potential negative outcomes, we highlight the tradeoffs between explainability and adversarial robustness and emphasize the importance of carefully balancing these two aspects. Our research focuses on understanding the vulnerabilities of these models and exploring ways to improve their resilience against adversarial attacks. Moreover, the insights obtained from this information can be utilized to develop models that are more robust and reliable, ultimately providing a means to foster civil discourse amidst a vast sea of potentially harmful interactions.

In terms of data ethics, our study utilizes public datasets and maintains the privacy and anonymity of the data subjects. This exploration does not involve collection of any new datasets. In the event of future data collection, we will strictly adhere to ethical guidelines.

Lastly, we have utilized LLMs to clean up the grammar and improve the readability of our paper. However, no content creation was performed using these tools.