# Analyzing Adversarial Strategies and Countermeasures for Cyberbullying Detection

Maddie Juarez[1], Eldor Abdukhamidov[2], Manuel Sandoval[1], Mujtaba Nazari[1], Deborah Hall[3], George K. Thiruvathukal[1], Tamer Abuhmed[2], Yasin N. Silva[1], and Mohammed Abuhamad[1]

[1] Loyola University Chicago, Chicago IL 60660, USA
[2] {mjuarez, msandovalmadrigal, mnazari, gkt, ysilva1, mabuhamad}@luc.edu
[3] Sungkyunkwan University, Suwon 2066, Republic of Korea
[4] {abdukhamidov, tamer}@skku.edu
[5] Arizona State University, Glendale AZ 85306, USA
[6] d.hall@asu.edu

**Abstract.** Cyberbullying on social networking sites has become more prevalent. Most cyberbullying detection models often lack consideration of adversarial threads, leaving them vulnerable. This study evaluates the resilience of text-based cyberbullying detection models, constrained by limited available datasets, against word-level substitutions and character-level perturbations. We consider well-established ML techniques with real-world data and more recent LLM-based approaches to uncover model weaknesses. The results reveal that adversarial attacks can significantly reduce detection accuracy, *e.g.,* most models are vulnerable to word- and character-level attacks with success rates up to 88% and 44%, respectively. We also find that LLM-based models such as CyberBERT are more resistant to both types of attack while maintaining strong detection performance. We show that model architecture and text vectorization choices significantly impact attack resistance and that adversarial training can help improve robustness, with tailored combinations of models and vectorizers showing the best results. These findings can guide the development of safer online platforms, as tailored strategies can make cyberbullying detection models more resilient and effective.

**Keywords:** Adversarial attack, Countermeasures, Cyberbullying

## 1 Introduction

Cyberbullying, the act of sending intentionally harmful and repetitive messages to others via electronic media, [14] has become a widespread problem with the rise of social media use. Given the robust link between cyberbullying and depression, substance abuse, and suicide [14] and the long-term consequences for victims, families, and communities, there is an urgent need for effective interventions and prevention strategies.
Researchers have proposed a range of algorithms designed to analyze online interactions and identify harmful behavior. While some approaches

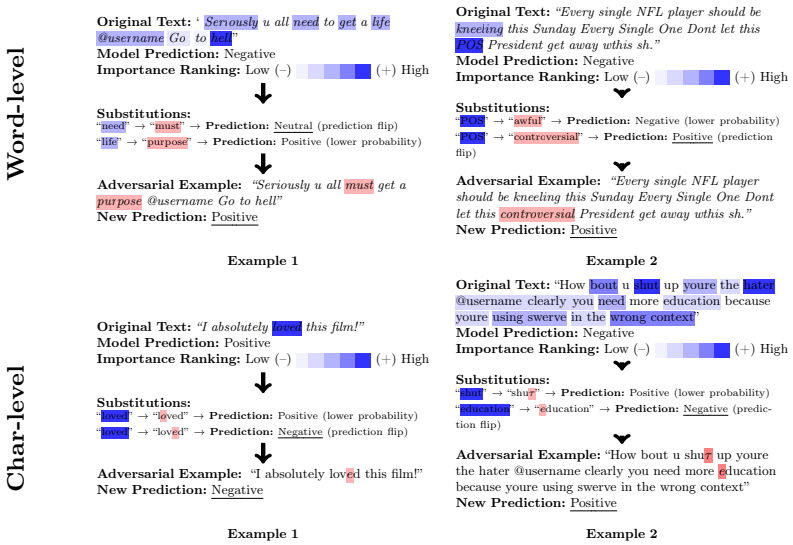


Fig. 1: Examples of word-level and char-level attacks.

incorporate multi-modal data (*e.g.,* images, user profiles, or network features)[10], most models are primarily text-based due to the limited and inconsistent access to additional metadata [21], a scarcity of large, expert-labeled datasets, and data access policies that limit the extraction and publication of contextual information. Although multiple cyberbullying detection models have been proposed, most of these models have not been designed considering mechanisms to effectively face adversarial attacks (*i.e.,* attacks that apply very small changes in the input to generate an incorrect classification outcome as shown in Figure 1). This study evaluates the robustness of cyberbullying detection against adversarial attacks and explores strategies that can make cyberbullying detection more resilient. Considering the mentioned data access challenges and the crucial importance of text data for cyberbullying detection, we restrict our analysis to textual perturbations. The main contributions of this paper are:

1. We perform a comprehensive vulnerability analysis of cyberbullying detection models against black-box adversarial attacks. We evaluate traditional ML models, deep learning approaches, and LLM-based models against word-level and character-level perturbations.
2. We evaluate cyberbullying detection methods using benchmark cyberbullying datasets from Instagram, Twitter, and Vine and evaluate the impact of adversarial attacks on model performance.
3. We discuss the effectiveness of adversarial training and other defense strategies to enhance the robustness of cyberbullying detection and provide recommendations for selecting model-vectorizer combinations to balance performance and resilience.

## 2 Related Work

**Cyberbullying Detection.** Many Natural Language Processing (NLP) approaches for detecting cyberbullying in social media have been proposed. These techniques feature well-known ML algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN, Random Forest (RF), and XGBoost [12,1]. Deep learning methods such as Long Short-Term Memory (LSTM) [12], Bidirectional Gated Recurrent Unit (Bi-GRU) [15,7], and Bidirectional LSTM (Bi-LSTM) [8] have also been proposed for bullying detection. In addition, researchers have proposed LLM-based models, such as CyberBERT [17], a fine-tuned version of BERT [5] for detecting cyberbullying. Researchers have also developed cyberbullying detection models that intrinsically depend on multi-modal data, such as images and temporal features, e.g., the model by Soni and Singh [20], HANCD [2] and XBully [4].

**Attacks Against Cyberbullying Detectors.** The robustness of cyberbullying detectors remains relatively unexplored. The only previous work in this area we are aware of [6], showed that 1) adversarial substitutions (replacing key terms with semantically similar alternatives) can significantly alter model predictions while retaining the original intent of the text, and 2) augmenting datasets with adversarial examples can enhance model robustness. We build on this initial evidence and differentiate from it as follows: **1)** whereas Emmery et al. only consider token-level perturbations, we consider token- and character-level perturbations to examine more granular and stealthier attacks, **2)** our analysis incorporates tokenized and vectorized representations, enabling nuanced evaluations of adversarial effects, **3)** instead of relying on calculated omission scores, we employ interpretation-guided techniques to target the most contextually significant tokens for substitution, and **4)** we evaluate models on dedicated cyberbullying datasets, enabling domain-specific analysis across the various models.

**Adversarial Attacks in Hate/Toxicity Detection.** Hosseini et al. [9] found that inserting punctuation marks or introducing minor misspellings in highly toxic terms can effectively bypass toxicity detectors. Other strategies include word obfuscation or polarity inversion [19] and nearest-neighbor substitution [11]. While cyberbullying shares some characteristics with hate speech and toxicity, it is distinguished by its repetitive and targeted nature, whereas hate or toxic content may involve isolated threats or broadly offensive language with minimal interaction [14].

## 3 Methodology

**Datasets** This study utilizes data from Instagram, X (formerly Twitter), and Vine (see table Table 1). Specifically, we employed 1) the Instagram dataset collected by [10], containing 2,219 social media sessions including post captions, comments, and labels indicating cyberbullying, 2) a Vine dataset [18] of 970 sessions, featuring videos, captions, comments, and

Table 1: Dataset statistics and model configuration.

| Model Type | Configuration | Data Representation |
|---|---|---|
| Traditional ML | SVM with RBF kernel<br>KNN with $k = 7$<br>Random Forest (1000 estimators)<br>XGBoost with decision trees | - TF-IDF (1000 features)<br>- TF-IDF with *bi*-grams<br>- TF-IDF with character-level $n$-grams<br>- Count vectorizer (1000 features) |
| DL | Bi-GRU (2 layers, 64 units each, embedding:128)<br>Bi-LSTM (2 layers, 64 units each, embedding:128) | Trainable embeddings (vocab size: 10,000)<br>Word2Vec (vocab size: 10,000) |
| Transformer | CyberBERT (fine-tuned with DistilBERT tokenizer) | DistilBERT embeddings (max sequence length: 30) |

| Dataset | #Sessions | #Bully | #Non-bully | #Comments/Tweets | Avg. Length |
|---|---|---|---|---|---|
| Instagram | 2,218 | 678 | 1,540 | 155,260 | 590.31 |
| Twitter | NA | 3,845 | 16,149 | 19,994 | 12.87 |
| Vine | 970 | 304 | 666 | 78,250 | 411.78 |

corresponding cyberbullying labels, and 3) a Twitter dataset [3] that contains 19,994 tweets labeled as bullying or not bullying. For model training, we split each dataset into a $k$-fold cross-validation ($k = 5$) scheme. When evaluating model performance, we use standard classification metrics: F1 Score, Recall, and Precision.

**Data Representation** We use the following methods to transform the textual data into numerical representations: 1) We apply two vectorization techniques: Term Frequency-Inverse Document Frequency (TF-IDF) and count-based vectorizer, using both character-level and word-level $n$-grams (ranging from 1 to 3 $n$-grams) to enable the models to process and extract meaningful features from the text data. For TF-IDF, we extract the top 1000 features based on term importance to focus on the most informative elements in the text. 2) We use pre-trained word embeddings such as Word2Vec [16] to capture semantic relationships between words and enhance the models' understanding of the text data. 3) We use pre-trained language models such as BERT to leverage contextual embeddings and improve the models' performance on complex text data.

**Machine Learning Models** We evaluate the robustness of the following cyberbullying models: SVM, NB, KNN, RF, XGBoost, LSTM, Bi-GRU, Bi-LSTM, and CyberBERT. We use the following configurations for the models: SVM with the Radial Basis Function (RBF) kernel, KNN with k=7, Random Forest with 1000 estimators, and XGBoost with decision trees. The non-LLM models are trained using *bi*-gram TF-IDF and TF vectorization, and $n$-grams character-based vectorization wth size $n$ = *1000*. We use a fined-tuned CyberBERT model with DistilBERT tokenizer and classification layers using a maximum sequence length of 30. Summary of the settings for models and data representation methods are shown in Table 1.

**World-level Attack** We set the threshold parameter $\tau = 0.1$, which determines whether a word or character substitution is accepted if it does not immediately flip the model's prediction. We assume a **black-box threat model** where the adversary only sees model outputs and probabilities. The adversary can **query** the model under limited bud-

Table 2: Performance of Cyberbullying Detection Models (the best performing results appear in bold and the second best are underlined).

| Method | Features | Instagram | | | Twitter | | | Vine | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** |
| XGBoost | TF-IDF 1000 | 0.73 | 0.83 | 0.65 | **0.76** | 0.79 | 0.74 | 0.60 | 0.74 | 0.55 |
| | TF-IDF *bi*-gram | 0.67 | 0.82 | 0.57 | 0.26 | 0.77 | 0.16 | 0.56 | 0.63 | 0.51 |
| | TF-IDF Char | 0.67 | 0.78 | 0.59 | 0.73 | 0.80 | 0.66 | 0.59 | 0.69 | 0.51 |
| | Count 1000 | 0.73 | 0.80 | 0.67 | 0.76 | 0.79 | 0.74 | 0.61 | 0.70 | 0.54 |
| RF | TF-IDF *bi*-gram | 0.65 | 0.88 | 0.55 | 0.34 | 0.24 | 0.55 | 0.59 | 0.74 | 0.49 |
| | TF-IDF Char | 0.51 | 0.87 | 0.36 | 0.34 | 0.22 | 0.72 | 0.46 | 0.80 | 0.3294 |
| | TF-IDF 1000 | 0.65 | **0.90** | 0.51 | 0.44 | 0.30 | 0.70 | 0.53 | 0.73 | 0.45 |
| | Count 1000 | 0.57 | 0.81 | 0.45 | 0.45 | 0.33 | 0.67 | 0.57 | 0.67 | 0.50 |
| SVM | TF-IDF *bi*-gram | 0.68 | 0.85 | 0.59 | 0.22 | 0.77 | 0.13 | 0.50 | **0.80** | 0.37 |
| | TF-IDF Char | 0.58 | 0.79 | 0.46 | 0.43 | 0.31 | 0.73 | 0.63 | 0.72 | 0.55 |
| | TF-IDF 1000 | 0.71 | 0.83 | 0.63 | 0.75 | 0.82 | 0.69 | 0.62 | 0.76 | 0.52 |
| | Count 1000 | 0.59 | 0.82 | 0.46 | 0.42 | 0.30 | 0.72 | 0.55 | 0.77 | 0.43 |
| NB | TF-IDF *bi*-gram | 0.51 | 0.38 | 0.78 | 0.55 | **0.88** | 0.40 | 0.50 | 0.37 | **0.90** |
| | TF-IDF Char | 0.36 | 0.24 | 0.77 | 0.44 | 0.37 | 0.58 | 0.55 | 0.43 | 0.73 |
| | TF-IDF 1000 | 0.62 | 0.43 | **0.81** | 0.48 | 0.44 | 0.52 | 0.60 | 0.48 | 0.72 |
| | Count 1000 | 0.43 | 0.41 | 0.45 | 0.47 | 0.44 | 0.49 | 0.42 | 0.52 | 0.36 |
| KNN | TF-IDF *bi*-gram | 0.45 | 0.51 | 0.39 | 0.37 | 0.53 | 0.29 | 0.56 | 0.48 | 0.67 |
| | TF-IDF Char | 0.63 | 0.61 | 0.66 | 0.38 | 0.57 | 0.28 | 0.65 | 0.54 | 0.84 |
| | TF-IDF 1000 | 0.56 | 0.58 | 0.54 | 0.21 | 0.44 | 0.14 | 0.59 | 0.46 | 0.84 |
| | Count 1000 | 0.43 | 0.69 | 0.31 | 0.34 | 0.47 | 0.27 | 0.57 | 0.73 | 0.48 |
| LSTM | Word2Vec | 0.34 | 0.45 | 0.28 | 0.50 | 0.49 | 0.50 | 0.51 | 0.48 | 0.57 |
| GRU | Embedding | 0.68 | 0.74 | 0.64 | 0.50 | 0.52 | 0.49 | 0.65 | 0.68 | 0.63 |
| CyberBERT | BERT | **0.75** | 0.76 | 0.73 | 0.75 | 0.71 | **0.79** | **0.74** | 0.75 | 0.73 |

get and substitute words with synonyms to preserve grammar and evade detection. The **goal** is to flip cyberbullying predictions—either to bypass moderation or falsely flag benign content. Following the TextFooler attack [13], we designed the attack to systematically identify and substitute influential words with semantically similar synonyms. This aims to generate adversarial examples that change the model's predictions while maintaining the original context and meaning of the text. The attack follows three main steps: 1) **Compute Importance Scores:** The algorithm calculates importance scores for each word by measuring how the model's prediction probability changes when each word is removed or masked. Words are then ranked by importance. 2) **Iterative Substitution:** Starting with the most important word, the algorithm finds semantically similar synonyms using pre-trained word embeddings (Word2Vec). Each synonym is tested—if it changes the model's prediction, it becomes the adversarial example. If a synonym only reduces prediction probability without changing the classification, the algorithm keeps track of the most impactful substitution and continues to the next word. 3) **Proceed to Next Important Word:** The process repeats until either the model's prediction changes or all important words have been processed. Examples of adversarial attacks are shown in Figure 1.

**Char-level Attack** We consider a practical attack that uses homoglyph substitutions (*i.e.,* visually similar characters with different Unicode encodings) against cyberbullying detectors. This attack presents key challenges: it can evade traditional token-based defenses while preserving

human readability, and it is simpler to implement compared to semantic transformations. Our approach combines homoglyph substitution with greedy word swaps guided by weighted saliency scores. The attack follows three key steps: 1) **Weighted Saliency Computation:** It calculates a weighted saliency score for each word measuring both its individual contribution to the model's prediction and its interactions with surrounding words. Higher scoring words are prioritized for perturbation. 2) **Greedy Homoglyph Substitution:** For words with higher saliency, it: a) generates visually similar glyphs for each character (*e.g.,* Latin 'a' → Cyrillic 'a'), b) tests each substitution by evaluating model predictions, c) accepts changes that flip prediction or maximally reduce confidence, and d) applies greedy selection for optimal character replacements. 3) **Iterate through Important Words:** The algorithm processes words in order of saliency score, preserves changes that reduce model confidence, and continues until prediction flips or all high-saliency words are tested. An example of the process is outlined in Figure 1.

## 4 Experiments and Results

Table 2 shows the baseline assessment of the cyberbullying detection performance of different models. The results show varying performance across different models and datasets. CyberBERT achieves the best F1 scores for Instagram (74.77%), Twitter (78.92%) and Vine (73.87%), while XGBoost with TF-IDF 1000 features performs best on Twitter (76.38%). SVM shows strong precision, particularly with TF-IDF *bi*-gram features reaching up to 85% on Instagram data. Deep learning approaches like LSTM and GRU demonstrate consistent but moderate performance across datasets.

**World-level Attack** Our findings highlight the importance of selecting appropriate model and vectorizer combinations for cyberbullying detection. Although effective in general classification tasks, GRU and SVM with char-level TF-IDF showed significant vulnerabilities to adversarial attacks. In contrast, models such as RF and KNN, especially when paired with vectorizers such as the TF vectorizer, demonstrated lower attack success rates, indicating greater resistance to adversarial exploitation. Table 3 presents the main word-level robustness evaluation results. GRU exhibited the highest attack success rates, 88% on the Instagram dataset, 87% on the Twitter dataset, and 88% on the Vine dataset. Similarly, LSTM showed relatively high success rates, with 73%, 59%, and 68% on Instagram, Twitter, and Vine datasets, respectively. CyberBERT showed moderate resistance to word-level attacks, with relatively low success rates across datasets.

**Char-level Attack** SVM with char-level TF-IDF shows notable vulnerability, with success rates of 13% on Instagram and 28% on Twitter. GRU also exhibits moderate susceptibility, reaching 42% on Twitter and 39% on Vine, indicating that models using finer-grained features are more easily manipulated, especially on shorter or simpler text. Random Forest demonstrates strong robustness, with consistently low success rates

Table 3: Word-level and Char-level Attack: Robustness evaluation with various vectorization techniques and datasets based on success rate, # of queries, and average perturbation size, *i.e.,* # words/characters changes.

| Models | Features | Word-level Attack | | | | | | | | | Char-level Attack | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Instagram | | | Twitter | | | Vine | | | Instagram | | | Twitter | | | Vine | | |
| | | Success Rate | # Query | Avg. chars | Success Rate | # Query | Avg. chars | Success Rate | # Query | Avg. chars | Success Rate | # Query | Avg. chars | Success Rate | # Query | Avg. chars | Success Rate | # Query | Avg. chars |
| XGBoost | TF-IDF 1000 | 0.41 | 4561 | 4.35 | 0.38 | 155 | 2.00 | 0.57 | 987 | 4.00 | 0.07 | 5183 | 3.18 | 0.19 | 62 | 3.99 | 0.16 | 1160 | 3.25 |
| | TF-IDF Char | 0.37 | 1291 | 7.48 | 0.61 | 53 | 4.15 | 0.59 | 334 | 7.32 | 0.09 | 4655 | 7.45 | 0.35 | 113 | 5.72 | 0.28 | 1236 | 5.51 |
| | TF 1000 | 0.45 | 4160 | 6.83 | 0.39 | 90 | 3.16 | 0.60 | 785 | 5.68 | 0.06 | 4421 | 2.83 | 0.16 | 78 | 4.63 | 0.21 | 1508 | 3.37 |
| | TF-IDF *bi*-gram | 0.19 | 4436 | 8.31 | 0.11 | 117 | 5.52 | 0.33 | 768 | 8.23 | 0.07 | 5948 | 3.83 | 0.07 | 73 | 4.11 | 0.10 | 1490 | 1.70 |
| SVM | TF-IDF 1000 | 0.19 | 10226 | 30.12 | 0.30 | 43 | 7.00 | 0.65 | 1675 | 56.75 | 0.04 | 3549 | 43.09 | 0.18 | 66 | 5.23 | 0.30 | 2401 | 25.64 |
| | TF-IDF Char | 0.82 | 1471 | 28.44 | 0.62 | 90 | 10.00 | 0.63 | 538 | 23.31 | 0.13 | 1643 | 47.79 | 0.28 | 83 | 8.29 | 0.24 | 1384 | 38.30 |
| | TF 1000 | 0.04 | 1994 | 21.00 | 0.23 | 70 | 4.00 | 0.29 | 1594 | 96.67 | 0.02 | 4382 | 76.83 | 0.13 | 53 | 4.78 | 0.10 | 1325 | 15.42 |
| | TF-IDF *bi*-gram | 0.06 | 8178 | 39.57 | 0.13 | 57 | 5.50 | 0.25 | 7178 | 157.48 | 0.04 | 3645 | 44.06 | 0.09 | 74 | 4.84 | 0.12 | 1630 | 70.12 |
| RF | TF-IDF 1000 | 0.03 | 2030 | 8.12 | 0.28 | 163 | 2.00 | 0.15 | 860 | 41.54 | 0.01 | 3266 | 6.25 | 0.19 | 38 | 5.28 | 0.08 | 1413 | 7.73 |
| | TF-IDF Char | 0.01 | 8272 | 6.70 | 0.34 | 123 | 3.00 | 0.18 | 794 | 24.36 | 0.00 | - | - | 0.13 | 65 | 5.68 | 0.05 | 877 | 7.00 |
| | TF 1000 | 0.02 | 2493 | 4.94 | 0.25 | 170 | 5.00 | 0.12 | 762 | 10.00 | 0.01 | 3527 | 8.45 | 0.17 | 57 | 4.27 | 0.08 | 1689 | 4.00 |
| | TF-IDF *bi*-gram | 0.04 | 1243 | 10.50 | 0.29 | 55 | 7.00 | 0.09 | 570 | 15.00 | 0.03 | 5944 | 8.45 | 0.17 | 57 | 4.27 | 0.08 | 1689 | 4.00 |
| NB | TF-IDF 1000 | 0.49 | 276 | 5.50 | 0.49 | 160 | 4.50 | 0.35 | 838 | 62.00 | **0.44** | 980 | 19.14 | 0.32 | 90 | 5.80 | 0.16 | 894 | 30.61 |
| | TF-IDF Char | 0.45 | 589 | 7.50 | 0.59 | 72 | 12.00 | 0.48 | 458 | 10.00 | 0.01 | 2077 | 77.65 | 0.31 | 53 | 6.82 | **0.41** | 616 | 95.21 |
| | TF 1000 | 0.18 | 1928 | 10.20 | 0.29 | 130 | 4.00 | 0.11 | 855 | 86.35 | 0.07 | 4460 | 22.43 | 0.13 | 92 | 5.72 | 0.05 | 2490 | 22.80 |
| | TF-IDF *bi*-gram | 0.01 | 2015 | 11.40 | 0.19 | 61 | 2.00 | 0.49 | 246 | 28.00 | 0.41 | 2357 | 2.00 | 0.13 | 68 | 6.42 | 0.39 | 497 | 39.26 |
| KNN | TF-IDF 1000 | 0.03 | 1670 | 3.40 | 0.14 | 72 | 3.50 | 0.10 | 299 | 3.37 | 0.003 | 3625 | 1.00 | 0.32 | 84 | 5.63 | 0.08 | 561 | 1.53 |
| | TF-IDF Char | 0.04 | 733 | 2.92 | 0.36 | 69 | 3.58 | 0.10 | 224 | 2.50 | 0.02 | 3831 | 1.20 | 0.32 | 82 | 5.16 | 0.03 | 481 | 1.00 |
| | TF 1000 | 0.01 | 629 | 2.00 | 0.30 | 72 | 2.98 | 0.06 | 674 | 2.18 | 0.003 | 3278 | 1.00 | 0.17 | 42 | 4.66 | 0.03 | 1102 | 1.00 |
| | TF-IDF *bi*-gram | 0.06 | 1341 | 3.53 | 0.16 | 78 | 4.12 | 0.25 | 533 | 3.31 | 0.02 | 2267 | 1.14 | **0.49** | 94 | 5.01 | 0.12 | 572 | 1.48 |
| LSTM | Word2Vec | 0.73 | 545 | 10.80 | 0.59 | 84 | 2.37 | 0.64 | 273 | 9.00 | – | – | – | – | – | – | – | – | – |
| GRU | Embeddings | **0.88** | 2568 | 14.26 | **0.87** | 55 | 6.00 | **0.88** | 985 | 18.05 | 0.08 | 766 | 4.64 | 0.42 | 77 | 5.89 | 0.39 | 1418 | 5.59 |
| CyberBERT | BERT | 0.12 | 1750 | 7.75 | 0.15 | 142 | 6.20 | 0.10 | 990 | 19.00 | – | – | – | 0.10 | 126 | 8.20 | 0.08 | 1400 | 43.75 |

across all datasets compared to other models. Even under heavy adversarial queries, the model maintains low vulnerability, particularly with the char-level TF-IDF vectorizer. Character-level perturbations lead to out-of-vocabulary (OOV) issues for models using pre-trained word embeddings. For LSTM using Word2Vec embeddings, all character-level perturbations result in OOV tokens across all datasets, making this attack method inapplicable (denoted by "–" in Table 3). Similarly, CyberBERT experiences OOV issues with character-level perturbations on the Instagram dataset, though it shows limited vulnerability on Twitter and Vine data where some perturbed tokens remain in the vocabulary.

## 5 Defenses and Countermeasures

**Suitable Synonyms for Word-level Attack** An attack may fail if no suitable synonyms exist. Fig. 2 shows examples of this phenomenon using samples from the Twitter dataset.

**Strong Contextual Dependencies** Maintaining coherence with word substitutions requires greater care in the presence of strong contextual dependencies, as seen in Fig. 3.

**Increasing Sophistication and Perturbation Budget**: Increasing both the sophistication of the altercations and the perturbation budget allows for more aggressive changes to the text. Unfortunately, as seen in Fig. 4, these changes come at the risk of destroying the intended meaning.

**Dataset Characteristics**: Char-level attacks are more effective against datasets with shorter and simpler text, as even small changes can significantly influence model predictions. For example, text originating from

**Original Text:** "My favorite kinds of Seahawk fans the any criticism of team youre a soulless ghoul who deserves sh"
**Adversarial Text:** "My favorite kinds of Seahawk followers the any critique of squad youre a heartless specter who deserves sh"
**Result:** The substitution of "fans" with "followers," "criticism" with "critique," and "ghoul" with "specter" reduced the model's confidence in classifying the tweet as *bullying* from 95% to 48%, thus flipping the prediction to *non-bullying*.

(a) Positive Outcome (Attack Successful)

**Original Text:** "Tomorrow on BraveNewWorlds I run my loveletter to Lazy Dungeon Masters and embracing the Fantastic"
**Attempted Adversarial Text:** "Tomorrow on BraveNewWorlds I run my ode to Slothful Dungeon Leaders and welcoming the Extraordinary"
**Result:** Despite the substitutions ("loveletter" with "ode," "Lazy," with "Slothful," and "Fantastic" with "Extraordinary"), the model continues to classify the text as *non-bullying* with 89% confidence.

(b) Negative Outcome (Attack Unsuccessful)

Fig. 2: Suitable Synonyms for Word-level Attack.

**Original Text:** "is a national treasure So many dickheads commenting BTL though"
**Adversarial Text:** "is a national treasure So many fools commenting below though"
**Result:** Substituting "dickheads" with "fools" and "BTL" with "below" diminishes the force of the statement. In turn, model confidence drops from 90% to 45% and the prediction flips from *bullying* to *non-bullying*.

(a) Contextual Dependency Disrupting Attack (Unsuccessful)

**Original Text:** "Just before they fought Tiamat at my DnD campaigns finale running for years lt"
**Attempted Adversarial Text:** "Just before they battled Tiamat at my DnD campaigns ending sprinting for years lt"
**Result:** Substituting "fought" with "battled," "finale" with "ending," and "running" with "sprinting" distorts the text, yet the model predicts *non-bullying* with 88% confidence.

(b) Preserved Contextual Dependency (Successful)

Fig. 3: Strong Contextual Dependencies.

Twitter is often shorter in length and less grammatically complex. Hence, such text is more susceptible to adversarial inputs at the character level. In contrast, text that originates from a context-rich environment and employs sophisticated grammatical structures may lead to models that are more resistant to char-level attacks.

## 6 Conclusion

In this study, we evaluated two types of adversarial attacks, word-level and char-level perturbations, against a set of known cyberbullying text classification models using datasets from Instagram, Twitter, and Vine. We observed that shallow models trained on TF-IDF vectorizers are the most susceptible to word-level attacks, with, for example, the SVM trained on char-level TF-IDF demonstrating an attack success rate of 82% and an average query length of 1471. Notably, though, this model

**Original Text:** "The implementation of this policy is problematic and may lead to widespread issues."
**Adversarial Text with Increased Perturbation Budget:** "The enactment of th1s regu1at1on is fl@wed and could engender pervasive trouble$."
**Word-level Substitutions:** "implementation" replaced with "enactment," "policy" replaced with "regulation," "problematic" replaced with "fl@wed."
**Character-level Substitutions:** "this" replaced with "th1s," "regulation" replaced with "regu1at1on," "issues" replaced with "trouble$."
**Increased Aggressiveness:** Multiple substitutions applied simultaneously, leading to a substantial change in readability and coherence.
**Result:** While the modified text successfully reduces a model's confidence in classifying the text as *bullying* (from 90% to 40%), the coherence of the text is significantly compromised.

(a) Increasing Sophisticated Techniques

**Original Text:** "The implementation of this policy is problematic and may lead to widespread issues."
**Adversarial Text with Increased Perturbation Budget:** "The enactment of th1s regu1at1on is fl@wed and could engender pervasive trouble$."
**Word-level Substitutions:** "implementation" replaced with "enactment," "policy" replaced with "regulation," "problematic" replaced with "fl@wed."
**Character-level Substitutions:** "this" replaced with "th1s," "regulation" replaced with "regu1at1on," "issues" replaced with "trouble$."
**Increased Aggressiveness:** Multiple substitutions applied simultaneously, leading to a substantial change in readability and coherence.
**Result:** While the modified text successfully reduces a model's confidence in classifying the text as *bullying* (from 90% to 40%), the coherence of the text is significantly compromised.

(b) Increasing Perturbation Budget

Fig. 4: Increasing Sophistication and Perturbation Budget.

was not the strongest cyberbullying detection model from the perspective of the F1 score. All models appeared resilient against char-level attacks, with the highest success rate being 44%, though, admittedly, this value is an outlier. We also observed that the Transformer-based Cyber-BERT model is quite resilient to word-level attacks, while also maintaining the highest F1 score on the cyberbullying detection task.

## 7 Acknowledgments

This work was supported by National Science Foundation Awards #2435164 and #2435165 and a Google Award for Inclusion Research.

## References

1. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
2. Cheng, L., Guo, R., Silva, Y.N., Hall, D., Liu, H.: Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. ACM/IMS Trans. Data Sci. **2**(2) (Apr 2021)
3. Cheng, L., Li, J., Silva, Y., Hall, D., Liu, H.: Pi-bully: Personalized cyberbullying detection with peer influence. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (2019)
4. Cheng, L., Li, J., Silva, Y.N., Hall, D.L., Liu, H.: Xbully: Cyberbullying detection within a multi-modal context. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
6. Emmery, C., Kádár, Á., Chrupała, G., Daelemans, W.: Cyberbullying classifiers are sensitive to model-agnostic perturbations. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022)

7. Fang, Y., Yang, S., Zhao, B., Huang, C.: Cyberbullying detection in social networks using bi-gru with self-attention mechanism. Information **12**(4) (2021)
8. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional lstm networks for improved phoneme classification and recognition. In: Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005 (2005)
9. Hosseini, H., Kannan, S., Zhang, B., Poovendran, R.: Deceiving google's perspective api built for detecting toxic comments (2017)
10. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: Detection of cyberbullying incidents on the instagram social network (2015)
11. Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z., Gu, H., Artan, N.S.: Adversarial text generation for google's perspective api. In: International Conference on Computational Science and Computational Intelligence (2018)
12. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer (2013), https://faculty.marshall.usc.edu/gareth-james/ISL/
13. Jin, D., Jin, Z., Zhou, J.T., Szolovits, P.: Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI conference on artificial intelligence (2020)
14. Kowalski, R., Giumetti, G., Schroeder, A., Lattanner, M.: Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth. Psychological bulletin **140** (02 2014)
15. Kumar, A., Sachdeva, N.: A bi-gru with attention and capsnet hybrid model for cyberbullying detection on social media. World Wide Web **25**(4), 1537–1550 (2022)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
17. Paul, S., Saha, S.: Cyberbert: Bert for cyberbullying identification. Multimedia Systems **28**, 1–8 (11 2020)
18. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., Mattson, S.A.: Careful what you share in six seconds: Detecting cyberbullying instances in vine. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2015)
19. Rodriguez, N., Rojas-Galeano, S.: Shielding google's language toxicity model against adversarial attacks (2018)
20. Soni, D., Singh, V.: Time reveals all wounds: Modeling temporal characteristics of cyberbullying. Proceedings of the International AAAI Conference on Web and Social Media **12** (06 2018)
21. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Montréal, Canada (Jun 2012)