Robust Hate Speech Detection via Mitigating Spurious Correlations

Kshitiz Tiwari
University of Arkansas
Fayetteville, AR, USA
ktiwari@uark.edu

Shuhan Yuan
Utah State University
Logan, UT, USA
shuhan.yuan@usu.edu

Lu Zhang
University of Arkansas
Fayetteville, AR, USA
lz006@uark.edu

Abstract

We develop a novel robust hate speech detection model that can defend against both word- and character-level adversarial attacks. We identify the essential factor that vanilla detection models are vulnerable to adversarial attacks is the spurious correlation between certain target words in the text and the prediction label. To mitigate such spurious correlation, we describe the process of hate speech detection by a causal graph. Then, we employ the causal strength to quantify the spurious correlation and formulate a regularized entropy loss function. We show that our method generalizes the backdoor adjustment technique in causal inference. Finally, the empirical evaluation shows the efficacy of our method. [1],

1 Introduction

Online social media bring people together and encourage people to share their thoughts freely. However, it also allows some users to misuse the platforms to promote the hateful language. As a result, hate speech, which "expresses hate or encourages violence towards a person or group based on characteristics such as race, religion, sex, or sexual orientation"[2], unfortunately becomes a common phenomenon on online social media. As a result, many online social media platforms such as Facebook and Twitter have policies prohibiting hate speech on their platforms. In order to prevent the spread of hate speech, programs have been deployed to automatically filter out hateful contents. However, in response to these programs, malicious users develop various approaches to evade detection, making hate speech very difficult to be detected by vanilla machine learning approaches. One of the common strategy is to deliberately revising texts, especially misspelling hate words, while preserving the intended meaning, such as typing the f-word as "fxxk". Some malicious users also replace racial slurs with other names, such as technology brands or products, to evade detection. Such strategy can be treated as the evasion attacks in the field of the adversarial attacks, where the adversary aims to evade detection by revising the malicious samples (Sun et al., 2020).

Research on defending against adversarial attacks in the text domain has been received significant attention in recent years (Wang et al., 2021a; Xu et al., 2020). However, how to make the hate speech detection model robust to malicious users is still under studied. Many existing adversarial defense methods assume that attackers replace the words in the original text by their synonyms in order to preserve semantic similarity (e.g., (Si et al., 2020; Ye et al., 2020)). However, in practice the malicious users may use the words with different semantic meanings for the word substitutions. For example, in the coded hate speech, the word "Google" may be used to represent "African-American" and "Skittles" may be used to indicate Muslim (Magu et al., 2017; Xu et al., 2022).

In this paper, we develop a novel robust hate speech detection model. We target the situation where a group of target words could be replaced with any words even with entire different semantic meanings. We identify the essential factor to defend such attacks as to capture the causation between the semantic meaning of input text and the label and remove the spurious correlation between them. To this end, we use causal graphs (Pearl, 2009) to describe the causal relationship among the semantic meaning of input text, the target words, and the label. The impact of the adversarial attack is modeled as the causal strength of the arrow between the target words and the label in the graph. We then formulate the learning problem by integrating the causal strength into a regularized entropy loss. Finally, we analyze the objective function and

[1] Code is available at: https://github.com/zthsk/CEBERT

[2] https://dictionary.cambridge.org/dictionary/english/hate-speech

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 51–56
November 20–23, 2022.

show that it generalizes the backdoor adjustment which is a technique widely used for removing spurious correlation in machine learning. The empirical evaluation shows that our method can defend against both word- and character-level attacks.

Related Work. Hate speech detection as a supervised text classification task has attracted a lot of attention in the natural language processing community (Badjatiya et al., 2017; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Rajamanickam et al., 2020; Tran et al., 2020; Mou et al., 2020). Vanilla hate speech detection techniques are vulnerable to adversarial attacks. Thus, several frameworks are proposed to achieve robustness on various adversarial attacks (Wang et al., 2021b) such as adversarial data augmentation (Si et al., 2020; Jin et al., 2020), adversarial training (Li and Qiu, 2020; Morris et al., 2020), and certified defenses (Ye et al., 2020; Zeng et al., 2021). Different from above works, we propose a causal graph-guided models and employ the causal strength to measure the impact of adversarial attacks. To the best of our knowledge, this is the first work that leverage causal modeling to tackle the challenge of adversarial attacks on hate speech detection.

2 Method

A hate speech detection model can be defined as a functional mapping from T to Y, where $t \in T$ is a set of input texts and $y \in Y$ is the target label set. In general, the output of the detection model is the softmax probability of predicting each class k, i.e., $f_k(t;\theta) = P(Y = y_k|t)$, where θ is the parameters of the model. We presume a given group of target words (usually hateful or sentiment words) denoted by H, and use X to indicate the remaining text excluding the words in H, i.e., $T = \langle X, H \rangle$. Adversarial examples are inputs to detection models with perturbations on H that purposely cause the model make mistakes.

2.1 Causal Graph for Hate Speech Detection

Causal graphs are widely used for representing causal relationships among variables (Pearl, 2009). A causal graph is a directed acyclic graph (DAG) $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ denotes a set of variables, and $\mathcal{E}$ indicates causal relationships.

We propose a causal graph for modeling the hate speech detection shown in Fig. 1. In this graph, in addition to X, H, Y, we also use I to indicate the hate intent from a user. As we cannot know the real intent of the user, we treat I as a hidden variable indicated by the dash circle. The causal graph can be explained as follows: if the user tends to share hateful content, he/she chooses the target words (which may be perturbed later) while expressing the hateful meaning in the rest part of the text. As a result, I is the parent of H and X, which are in turn the parents of Y. For example, given a text T, e.g., "We don't want more [religious group] in this country. Enough is enough with those MAGGOTS.", H is the word "MAGGOTS" while X indicates the remaining text.

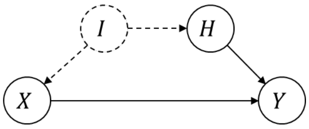

Figure 1: The causal graph for hate speech detection.

Based on the causal graph, we identify one major reason that vanilla detection models are not robust to adversarial attacks: the detection models make predictions based on both the semantic meanings of texts and the spurious correlation between X and Y via H (i.e., $X \leftarrow I \rightarrow H \rightarrow Y$) that significantly relates to the occurrence of the target words. When the target works, like the f-word, are strongly correlated with the hate label in the training dataset, the model trained on such data may easily make predictions based on the occurrence of the target words without considering the meanings of entire texts. Therefore, once the adversarial attacks that remove such correlations are conducted, the detection model is easy to be fooled.

2.2 Causal Strength for Measuring Spurious Correlation

In order to make the detection model robust to any perturbation, one needs to prevent the model from learning the spurious correlation. To this end, we propose to penalize the causal influence of H on Y during the training so that the spurious correlation can be blocked. Inferring causal influences of input on predictions is a challenging task in machine learning. In this paper, we advocate the use of the causal strength proposed in (Janzing et al., 2013), the idea of which is to measure the impact of an intervention that removes certain arrows in the causal graph. This definition naturally aligns with our context where we want to measure the impact of removing the correlation between the target words and the hate labels by modifying the target words, i.e., the causal strength of the arrow $H \to Y$.

Symbolically, denote the causal strength of $H \to Y$ by $\mathfrak{C}_{H\to Y}$. Quantifying $\mathfrak{C}_{H\to Y}$ requires to consider the conditional distribution of Y should we cut the arrow $H \to Y$. This distribution, which is referred to as the "post-cutting" distribution in (Janzing et al., 2013), is given by

$$P_{H\to Y}(y|x) = \sum_{h'\in H} P(y|x,h')P(h'). \quad (1)$$

Denote by P and $P_{H\to Y}$ the factual joint distribution and the "post-cutting" joint distribution respectively. Then, the causal strength $\mathfrak{C}_{H\to Y}$ is given by the Kullback–Leibler divergence $D[P||P_{H\to Y}]$, i.e., $\mathfrak{C}_{H\to Y} =$

$$\begin{aligned} &D[P||P_{H\to Y}] = D[P(Y|X,H)||P_{H\to Y}(Y|X)] \\ &= \sum_{x,h,y} P(x,h,y)\log\frac{P(y|x,h)}{\sum_{h'}P(y|x,h')P(h')}, \end{aligned} \quad (2)$$

where the second equality is due to factorization.

2.3 Problem Formulation

Since the causal strength measures the influence of the word substitution, our problem becomes to penalize the causal strength in the training. In order to integrate the causal strength into the objective function, we rewrite Eq. (2) according to the quotient rule for logarithms as follows.

$$\begin{aligned} \mathfrak{C}_{H\to Y} &= \sum_{x,h,y} P(x,h,y)\log P(y|x,h) \\ &- \sum_{x,h,y} P(x,h,y)\log\sum_{h'} P(y|x,h')P(h'). \end{aligned} \quad (3)$$

For the first term of Eq. (3), note that if we replace $P(y|x,h)$ with the parameterized function of the detection model and estimate $P(x,h,y)$ with the empirical distribution from the data, it can be reformulated as the same form as the cross-entropy loss with the reversed sign. We denote it by $-\mathcal{L}_{CE}$, i.e.,

$$-\mathcal{L}_{CE} = \frac{1}{N}\sum_j\sum_k y_k^{(j)}\log f_k(t^{(j)}),$$

where N is the number of text in the data, j indicates the j-th text, and k is the class index. We similarly reformulate the second term of Eq. (3), denoted by $\mathcal{L}_I$, i.e.,

$$\mathcal{L}_I = -\frac{1}{N}\sum_j\sum_k y_k^{(j)}\log\sum_{h'} f_k(t^{(j)})P(h').$$

Finally, by adding the causal strength as a regularization term into the cross-entropy loss, we obtain the regularized cross-entropy loss as follows.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathfrak{C}_{H\to Y} = (1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_I, \quad (4)$$

where $\lambda \in [0,1]$ is the coefficient for balancing the model utility and the model robustness.

2.4 Connection to Backdoor Adjustment

We further analyze the meaning of the term $\mathcal{L}_I$ in Eq. (4). As mentioned earlier, the reason that causes the traditional detection model to be vulnerable to adversarial attacks is the spurious correlation between X and Y. The backdoor adjustment is a classic technique for removing the spurious correlation (Pearl, 2009). It has been applied to various tasks like image captioning (Yang et al., 2021) and question answering (Qi et al., 2020) to improve the model robustness. In our context, this idea means to use the interventional distribution $P(Y|do(X))$ instead of the actual distributions $P(Y|X,H)$ or $P(Y|X)$ for predicting the label, where $do(\cdot)$ is the do-operator (Pearl, 2009) in Pearl's structural causal model that performs an intervention on the input variable (i.e., X in our case).

By applying the backdoor adjustment based on the causal graph Fig. 1, the interventional distribution $P(Y|do(X))$ is computed as

$$\begin{aligned} P(y|do(x)) &= \sum_{h',i} P(i)P(h'|i)P(y|x,h') \\ &= \sum_{h'} P(h')P(y|x,h'). \end{aligned} \quad (5)$$

Comparing Eqs. (1) and (5), we see an expected coincidence in the two formulas. This is because both the "arrow cutting" and the backdoor adjustment break the path $X \leftarrow I \to H \to Y$. The issue of directly using the interventional distribution $P(Y|do(X))$ for the prediction is that the model utility depends on how close $P(Y|do(X))$ is to the actual distribution, which cannot be controlled by the user. Thus, our loss formulation Eq. (4) can be considered as a generalization to the backdoor adjustment-based approaches, which is grounded on the causal strength theorem.

2.5 Practical Considerations

In Eq. (1), there is a summation over all the possible target words. Since target words are usually sentiment words, in this paper we propose to build a sentiment lexicon that includes the commonly

Model	Clean Dataset				
	Accuracy	Precision	Recall	Pos. Class F1	Macro F1
BERT Base	0.909±0.002	0.945±0.002	0.944±0.003	0.945±0.001	0.840±0.000
hateBERT	0.910±0.001	0.948±0.001	0.942±0.001	0.945±0.001	0.846±0.005
RANMASK	0.908±0.006	0.923±0.046	**0.945±0.011**	0.944±0.003	0.840±0.016
TAVAT	**0.916±0.002**	**0.966±0.006**	0.931±0.007	**0.948±0.001**	**0.864±0.005**
MIXADA	0.912±0.003	0.954±0.009	0.939±0.008	0.946±0.002	0.854±0.009
CEBERT	0.876±0.002	0.915±0.002	0.936±0.002	0.925±0.001	0.774±0.005

Table 1: Results on the clean test dataset.

Model	Replaced Dataset				
	Accuracy	Precision	Recall	Pos. Class F1	Macro F1
BERT Base	0.696±0.004	0.887±0.004	0.723±0.007	0.797±0.003	0.596±0.005
hateBERT	0.703±0.009	0.895±0.004	0.724±0.010	0.801±0.007	0.606±0.011
RANMASK	0.698±0.027	0.882±0.011	0.733±0.047	0.800±0.025	0.592±0.016
TAVAT	0.676±0.038	0.902±0.007	0.682±0.057	0.775±0.036	0.594±0.024
MIXADA	0.696±0.022	0.895±0.007	0.716±0.035	0.795±0.020	0.604±0.015
CEBERT	**0.859±0.002**	**0.909±0.001**	**0.922±0.002**	**0.915±0.001**	**0.750±0.000**

Table 2: Results on the replaced test dataset.

Model	Misspelled Dataset				
	Accuracy	Precision	Recall	Pos. Class F1	Macro F1
BERT Base	0.732±0.005	0.924±0.005	0.729±0.019	0.802±0.038	0.654±0.005
hateBERT	0.737±0.031	0.939±0.003	0.728±0.038	0.820±0.026	0.666±0.027
RANMASK	0.723±0.034	0.925±0.019	0.726±0.056	0.811±0.031	0.642±0.023
TAVAT	0.727±0.039	**0.948±0.012**	0.709±0.060	0.810±0.036	0.660±0.027
MIXADA	0.726±0.007	0.938±0.007	0.716±0.015	0.812±0.007	0.656±0.005
CEBERT	**0.860±0.002**	0.909±0.002	**0.922±0.004**	**0.916±0.001**	**0.752±0.004**

Table 3: Results on the misspelled test dataset.

used sentiment words. Note that the words in the lexicon do not need to be synonyms of particular sentiment words and can include both hate and non-hate words. In our experiments, we construct the lexicon based on the hate word vocabulary provided by Ahn[3] and the positive word vocabulary provided by Parade[4].

3 Empirical Evaluation

3.1 Experimental Setting

We first build a list L of target words based on Ahn and Parade that contains 446 hate words and 126 non-hate words. We then randomly select m words from the list as our sentiment lexicon H. The default value of m is 16 in the experiments.

We curate a dataset by combining three dataset that are frequently used for hate speech detection: the OLID dataset (Zampieri et al., 2019), the White Supremacy Forum (De Gibert et al., 2018), and the AHSD dataset (Davidson et al., 2017). The combined dataset is then pre-processed by removing texts that do not contain any word in the list L. The resulting dataset contains 27368 texts among which 4818 texts are regular and 22550 texts are hate. It is then randomly split into training and test set by the ratio 4:1. Each experiment is repeated five times using different random seeds.

We consider five baselines in the experiments: the base BERT and HateBERT (Caselli et al., 2021) are vanilla detection models; MixADA (Si et al., 2021) is an adversarial data augmentation method; TAVAT (Li and Qiu, 2021) is an adversarial training method; and RanMask (Zeng et al., 2021) is a certified defense method.

To evaluate the robustness of all models, we use three different versions of the test dataset: the clean version, the word-level attack version where each word from the texts present in the list L is randomly replaced by one of the words in L, and the character-level attack version where each word in L is replaced by a misspelled version.

Our model uses the pre-trained BERT as the base model which is then fine-tuned by minimizing Eq. (4) on our training data. By default $\lambda = 0.5$. The prior probability $P(h')$ for a target word h' is calculated by dividing the total occurrence of h' in the training data by the total occurrence of all the words in L in the training data. We refer to our

[3] https://www.cs.cmu.edu/~biglou/resources/

[4] https://parade.com/1241177/marynliles/positive-words/

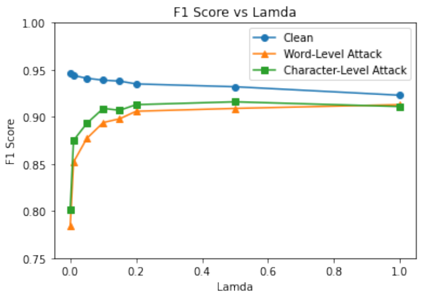

Figure 2: Pos. class F1 versus λ in Eq. (4) on different datasets.

model the **CEBERT**.

3.2 Experimental Results

Robust Hate Speech Detection. We first evaluate the performance of all models on three test datasets in terms of accuracy, precision, recall and F1 scores of the positive (i.e., hate) class as well as the Macro F1. The mean and standard deviation of five runs are shown in Table 1. As can be seen, the base BERT model produces good accuracy and F1 on the clean data but the worst results on the misspelled dataset. Other baselines improve the performance on the perturbed datasets, but the improvements are limited. CEBERT, on the other hand, trades of the performance on the clean data for the robustness and achieves the best performance on the perturbed datasets with a large margin compared with baselines.

Sensitivity Analysis. We also evaluate the influence of λ in Eq. (4) on CEBERT that balances $\mathcal{L}_{CE}$ and $\mathcal{L}_I$. We can observe from Fig. 2 that only using the $\mathcal{L}_I$ loss ($\lambda = 1$) to fine-tune the BERT model can achieve the best performance on the perturbed datasets, but the performance on the clean dataset becomes slightly worse. On the other hand, a small value of λ in range between 0.1 and 0.2 can produce a balanced performance.

4 Conclusions

We developed a robust hate speech detection model by leveraging the causal inference to mitigate spurious correlations. The experiment results show that our model can achieve better performance under both word- and character-level attacks compared with other baselines.

Acknowledgement

This work was supported in part by NSF 1946391.

Ethical Considerations

In this paper, we have improved the robustness of hate speech detection. One limitation of our proposed method is it assumes that we are given a list of target words that could be manipulated. If the list does not contain all target words, then the performance of our method may be lower than expected.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2013. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. 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*, volume 34, pages 8018–8025.

Linyang Li and Xipeng Qiu. 2020. Tavat: Token-aware virtual adversarial training for language understanding. *arXiv preprint arXiv:2004.14543*.

Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8410–8418.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 608–611.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *arXiv:2005.05909 [cs]*.

Guanyi Mou, Pengyi Ye, and Kyumin Lee. 2020. SWE2: SubWord Enriched and Significant Word Emphasized Framework for Hate Speech Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1145–1154, New York, NY, USA. Association for Computing Machinery.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *ACL*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. *arXiv preprint arXiv:2012.15699*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. In *Findings of ACL*.

Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S. Yu, Lifang He, and Bo Li. 2020. Adversarial Attack and Defense on Graph Data: A Survey. *arXiv:1812.10528 [cs]*.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Serim Park. 2020. HABERTOR: An Efficient and Effective Deep Hatespeech Detector. In *EMNLP*.

Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2021a. Towards a robust deep neural network against adversarial texts: A survey. *ieee transactions on knowledge and data engineering*.

Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2021b. Towards a Robust Deep Neural Network in Texts: A Survey. *arXiv:1902.07285 [cs]*.

Depeng Xu, Shuhan Yuan, Yueyang Wang, Angela Uchechukwu Nwude, Lu Zhang, Anna Zajicek, and Xintao Wu. 2022. Coded hate speech detection via contextual information. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 93–105. Springer.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv preprint arXiv:2005.14424*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. 2021. Certified robustness to text adversarial attacks by randomized [mask]. *arXiv preprint arXiv:2105.03743*.