Deceiving Google's Perspective API Built for Detecting Toxic Comments

Hossein Hosseini, Sreeram Kannan, Baosen Zhang and Radha Poovendran
Network Security Lab (NSL), Department of Electrical Engineering, University of Washington, Seattle, WA
Email: {hosseinh, ksreeram, zhangbao, rp3}@uw.edu

***Abstract*—Social media platforms provide an environment where people can freely engage in discussions. Unfortunately, they also enable several problems, such as online harassment. Recently, Google and Jigsaw started a project called Perspective, which uses machine learning to automatically detect toxic language. A demonstration website has been also launched, which allows anyone to type a phrase in the interface and instantaneously see the toxicity score [1].**

In this paper, we propose an attack on the Perspective toxic detection system based on the adversarial examples. We show that an adversary can subtly modify a highly toxic phrase in a way that the system assigns significantly lower toxicity score to it. We apply the attack on the sample phrases provided in the Perspective website and show that we can consistently reduce the toxicity scores to the level of the non-toxic phrases. The existence of adversarial examples is very harmful for toxic detection system and seriously undermines its usability.

I. Introduction

Social media platforms provide an environment where people can learn about the trends and news, freely share their opinions and engage in discussions. Unfortunately, the lack of a moderating entity in these platforms has caused several problems, ranging from the wide spread of fake news to online harassment [2]. Due to the growing concern about the impact of online harassment on the people's experience of the Internet, many platforms are taking steps to enhance the safety of the online environments [3], [4].

Some of the platforms employ approaches such as crowdsourcing (upvotes/downvotes), turning off comments or manual moderation to mitigate the effect of the inappropriate contents [5]. These approaches however are inefficient and not scalable. As a result, there has been many calls for researchers to develop methods to automatically detect the abusive or toxic context in the real time [6].

Recent advances in machine learning have transformed many domains such as computer vision [7], speech recognition [8], and language processing [9]. Many researchers have explored using machine learning to tackle the problem of online harassment. Recently, Google and Jigsaw launched a project called Perspective [1], which uses machine learning to automatically detect online insults, harassment, and abusive speech. The system intends to bring Conversation AI to help with providing a safe environment for online discussions [10].

Perspective is an API that enables the developers to use the toxic detector running on Google's servers, for identifying harassment and abuse on social media or more efficiently filtering invective from the comments on a news website. Jigsaw has partnered with online communities and publishers, such as Wikipedia [3] and The New York Times [11], to implement the toxicity measurement system.

Recently, a demonstration website has been launched, which allows anyone to type a phrase in the Perspective's interface and instantaneously see how it rates on the "toxicity" scale [1]. The Perspective website also states that the experiments, models and research data are open sourced in order to explore the strengths and weaknesses of using machine learning as a tool for online discussion.

The implicit assumption of learning models is that they will be deployed in benign settings. However, many works have pointed out their vulnerability in adversarial scenarios [12]–[14]. One type of the vulnerabilities of machine learning algorithms is that an adversary can change the algorithm output by subtly perturbing the input, often unnoticeable by humans. Such inputs are called *adversarial examples* [15], and have been shown to be effective against different machine learning algorithms even when the adversary has only a black-box access to the victim model [16].

In this paper, we demonstrate the vulnerability of the recently-proposed Google's Perspective system against the adversarial examples. In the text classification task of the Perspective, adversarial examples can be defined as modified texts which contain the same highly abusive language as the original text, yet receive a significantly lower toxicity score from the learning model. Through different experiments, we show that an adversary can deceive the system by misspelling the abusive words or by adding punctuations between their letters. The existence of adversarial examples is very harmful for toxic detector systems and seriously undermines their usability, especially since these systems are likely to be employed in adversarial settings. We conclude the paper by proposing some countermeasures to the proposed attack.

II. Background

A. Brief Description of Google's Perspective API

Perspective is an API created by Jigsaw and Google's Counter Abuse Technology team in Conversation-AI. Conversation AI is a collaborative research effort exploring ML as a tool for better discussions online [17]. The API uses machine

This work was supported by ONR grants N00014-14-1-0029 and N00014-16-1-2710 and ARO grant W911NF-16-1-0485.

TABLE I: Demosntration of the Attack on the `Perspective` Toxic Detection System . All phrases in the first column of the table are chosen from the examples provided by the `Perspective` website [1].

Original Phrase (Toxicity Score)	**Modified Phrase (Toxicity Score)**
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%)	Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%)
They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%)	They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%)
They are liberal **idiots** who are **uneducated** (90%)	They are liberal **i.diots** who are **un.educated** (15%)
idiots. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%)	**idiiots**. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%)
They are **stupid** and **ignorant** with no class (91%)	They are **st.upid** and **ig.norant** with no class (11%)
It's **stupid** and wrong (89%)	It's **stuipd** and wrong (17%)
If they voted for Hilary they are **idiots** (90%)	If they voted for Hilary they are **id.iots** (12%)
Anyone who voted for Trump is a **moron** (80%)	Anyone who voted for Trump is a **mo.ron** (13%)
Screw you trump supporters (79%)	**S c r e w** you trump supporters (17%)

learning models to score the toxicity of an input text, where toxic is defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion."

Google and Jigsaw developed the measurement tool by taking millions of comments from different publishers and then asking panels of ten people to rate the comments on a scale from "very toxic" to "very healthy" contribution. The resulting judgments provided a large set of training examples for the machine learning model.

Jigsaw has partnered with online communities and publishers to implement the toxicity measurement system. Wikipedia use it to perform a study of its editorial discussion pages [3] and The New York Times is planning to use it as a first pass of all its comments, automatically flagging abusive ones for its team of human moderators [11]. The API outputs the scores in real-time, so that publishers can integrate it into their website to show toxicity ratings to commenters even during the typing [5].

B. Adversarial Examples for Learning Systems

Machine learning models are generally designed to yield the best performance on clean data and in benign settings. As a result, they are subject to attacks in adversarial scenarios [12]–[14]. One type of the vulnerabilities of the machine learning algorithms is that an adversary can change the algorithm prediction score by perturbing the input slightly, often unnoticeable by humans. Such inputs are called *adversarial examples* [15], and have been shown to be effective against different machine learning algorithms even if the adversary has only a black-box access to the victim model [16].

Adversarial examples have been applied to models for different tasks, such as images classification [15], [18], [19], music content analysis [20] and malware classification [21]. In this work, we consider the problem of generating adversarial examples for text classification. In the context of scoring the toxicity, adversarial examples can be defined as modified texts which contain the same highly abusive language as the original one, yet receive a significantly lower toxicity score by the model.

In a similar work, the authors proposed a method for gender obfuscating in social media writing [22]. The proposed method modifies the text such that the writer is classified as a certain target gender, under limited knowledge of the classifier and while preserving the text's fluency and meaning. The modified texts are not required to be adversarial, i.e., a human may also classify it as the target gender. In contrast, in the application of toxic text detection, the adversary intends to deceive the classifier, while *maintaining the abusive content of the text*.

III. The Proposed Attacks

Recently, a website has been launched for `Perspective` demonstration, which allows anyone to type a phrase in the interface and instantaneously receive the toxicity score [1]. The website provides three categories of topics "that are often difficult to discuss online". The categories are 1) Climate Change, 2) Brexit and 3) US Election.

In this section, we demonstrate an attack on the `Perspective` toxic detection system, based on the adversarial examples. In particular, we show that an adversary can subtly modify a toxic phrase such that the model will output a very low toxicity score for the modified phrase. The attack setting is as follows. The adversary possesses a phrase with a toxic content and tries different perturbations on the words, until she succeeds with significantly reducing the confidence of the model that the phrase is toxic. Note that the adversary does not have access to the model or training data, and can only query the model and get the toxicity score.

Table I demonstrates the attack on sample phrases provided by the `Perspective` website. The first column represents the original phrases along with the toxicity scores and the second column provides the adversarially modified phrases and their corresponding toxicity scores. For better demonstration of the attack, we chose phrases with different toxic words and also introduced different types of errors, rather than searching for the best error type that would potentially yield lower toxicity score. The boldface words are the toxic words that the adversary has modified. The modifications are adding a dot between two letters, adding spaces between all letters and misspelling the word (repeating one letter twice or swapping

TABLE II: Demosntration of False Alarm on the Perspective Toxic Detection System . All phrases in the first column of the table are chosen from the examples provided by the Perspective website [1]

Original Phrase (Toxicity Score)	**Modified Phrase (Toxicity Score)**
Climate change is happening and it's not changing in our favor. If you think differently you're an idiot (84%)	Climate change is happening and it's not changing in our favor. If you think differently you're **not** an idiot (73%)
They're stupid, it's getting warmer, we should enjoy it while it lasts (86%)	They're **not** stupid, it's getting warmer, we should enjoy it while it lasts (74%)
They are liberal idiots who are uneducated. (90%)	They are **not** liberal idiots who are uneducated. (83%)
idiots. backward thinking people. nationalists. not accepting facts. susceptible to lies. (80%)	**not** idiots. **not** backward thinking people. **not** nationalists. accepting facts. **not** susceptible to lies. (74%)
They are stupid and ignorant with no class (91%)	They are **not** stupid and ignorant with no class (84%)
It's stupid and wrong (89%)	It's **not** stupid and wrong (83%)
If they voted for Hilary they are idiots (90%)	If they voted for Hilary they are **not** idiots (81%)
Anyone who voted for Trump is a moron (80%)	Anyone who voted for Trump is **not** a moron (65%)
Screw you trump supporters (79%)	**Will not** screw you trump supporters (68%)

two letters). As can be seen, we can consistently reduce the toxicity score to the level of the benign phrases by subtly modifying the toxic words.

Moreover, we observed that the adversarial perturbations *transfer* among different phrases, i.e., if a certain modification to a word reduces the toxicity score of a phrase, the same modification to the word will reduce the toxicity score also in another phrase. Using this property, an adversary can form a dictionary of the adversarial perturbations for every word and significantly simplify the attack process.

Through the experiments, we made the following observations:

- Susceptibility to false alarm: we observed that the Perspective system also wrongly assigns high toxicity scores to the apparently benign phrases. Table II demonstrates the false alarm on the same sample phrases of Table I. The first column represents the original phrases along with the toxicity scores and the second column provides the negated phrases and the corresponding toxicity scores. The boldface words are added to the second column. As can be seen, the system consistently fails to capture the inherent semantic of the modified phrases and wrongly assigns high toxicity scores to them.
- Robustness to random misspellings: we observed that the system assigns 34% toxicity score to most of the misspelled and random words. Also, it is somewhat robust to phrases which contain randomly perturbed toxic words.
- Vulnerability to poisoning attack: The Perspective interface allows the users to provide a feedback on the toxicity of the phrases, suggesting that the the learning algorithm updates itself using the new data. This can expose the system to poisoning attacks, where an adversary modifies the training data (in this case, the labels) such that it will assign low toxicity score to certain phrases.

IV. Open Problems in Defense Methods

The developers of Perspective have mentioned that the system is in the early days of research and development, and that the experiments, models, and research data are published to explore the strengths and weaknesses of using machine learning as a tool for online discussion.

In section III, we showed the vulnerability of the Perspective system against the adversarial examples. Scoring the semantic toxicity of a phrase is clearly a very challenging task. In this following, we briefly review some of the possible approaches for improving the robustness of the toxic detection systems:

- Adversarial Training: In this approach, during the training phase, we generate the adversarial examples and train the model to assign the original label to them [18]. In the context of toxic detection systems, we need to include different modified versions of the toxic words into the training data. While this approach may improve the robustness of the system against the adversarial examples, it does not seem practical to train the model on all variants of every word.
- Spell checking: Many of the adversarial examples can be detected by first applying a spell checking filter before the toxic detection system. This approach may however increase the false alarm.
- Block suspicious users for a period of time: The adversary needs to try different error patterns to finally evade the toxic detection system. Once a user fails to pass the threshold for a number of times, the system can block her for a while. This approach can force the users to less often use toxic language.

V. Conclusion

In this paper, we studied the feasibility of attacking the recently-released Google's Perspective API built for detecting toxic comments. We showed that the system can be deceived by slightly perturbing the abusive phrase to receive very low toxicity scores, while preserving the intended meaning. We also showed that the system has high false alarm rate in scoring high toxicity scores to benign phrases. We provided detailed examples for the studied cases. Our future work includes development of countermeasures against such attacks.

Disclaimer: The phrases used in Tables I and II are chosen from the examples provided in the Perspective website [1] for the purpose of demonstrating the results and do not represent the view or opinions of the authors or sponsoring agencies.

References

[1] https://www.perspectiveapi.com/.

[2] M. Duggan, *Online harassment*. Pew Research Center, 2014.

[3] https://meta.wikimedia.org/wiki/Research:Detox.

[4] https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html.

[5] https://www.wired.com/2017/02/googles-troll-fighting-ai-now-belongs-world/.

[6] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," *arXiv preprint arXiv:1610.08914*, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

[10] https://jigsaw.google.com/.

[11] http://www.nytco.com/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/.

[12] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pp. 16–25, ACM, 2006.

[13] M. Barreno, B. Nelson, A. D. Joseph, and J. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[14] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43–58, ACM, 2011.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[16] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697*, 2016.

[17] https://conversationai.github.io/.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[19] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

[20] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015.

[21] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," *arXiv preprint arXiv:1606.04435*, 2016.

[22] S. Reddy, M. Wellesley, K. Knight, and C. Marina del Rey, "Obfuscating gender in social media writing," *NLP+ CSS 2016*, p. 17, 2016.