Adversarial Text Generation for Google's Perspective

REU fellows: Stephan Brown[4], Edwin Jain[1], Jeffrey Chen[3], Mohammad Baidas[4], Erin Neaton[2]
Faculty Mentors: Drs. N. Sertac Artan[4], Huanying Gu[4], Ziqian Dong[4]
Affiliation: [1.] Tufts University, [2.] University of Michigan, [3.] University of California Berkley, [4.] School of Engineering and Computing Science, NYIT
Emails: nartan@nyit.edu, hgu03@nyit.edu, ziqian.dong@nyit.edu

ABSTRACT

With the preponderance of harassment and abuse, social media platforms and online discussion platforms seek to curb toxic comments. Google's Perspective seeks to help platforms classify toxic comments. We have created a pipeline to modify toxic comments to fool and evade Perspective. This pipeline uses existing adversarial machine learning attacks to find the optimal perturbation which will fool the model. Since these attacks typically target images, as opposed to discrete text data, we include a process to generate text candidates from perturbed features and select candidates to retain syntactic similarity. With just 10,000 queries, changing three words in each comment fools Perspective 25% of the time, suggesting that building a surrogate model may not require many queries. We hope classifiers can improve their robustness via adversarial training.

BACKGROUND

Users on social media platforms are often subject to online abuse, trolling and harassment. To tackle this problem, Google launched a project called Perspective which uses machine learning to perform text classification and rate the "toxicity" of comments.

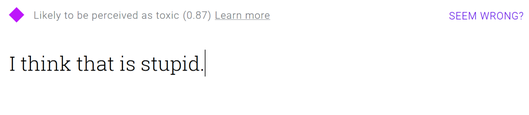


Fig. 1: Example of Comment Classification with Perspective

We have designed a targeted attack scheme to make Perspective misclassify toxic comments as clean.

Major Challenges

Text is discrete, not continuous.

Technically, pixels in a valid image must be integer values, but the discretization process is as simple as rounding. Discretizing some random vector into a word, on the other hand, is more difficult.

Perturbations to text can alter meaning.

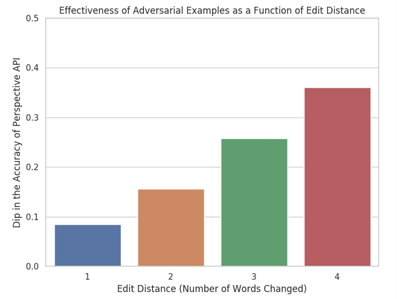
Slightly changing all of the pixels in an image is unlikely to change its label for humans, but this is obviously not the case for text. Furthermore, whereas one-pixel perturbations are almost guaranteed to be imperceptible, changing one word can actually alter sentiment.

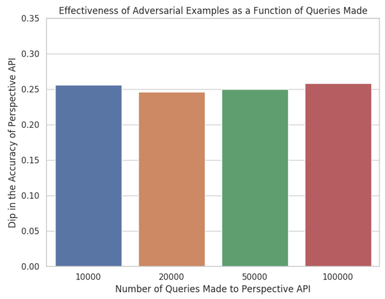
Language is dynamic.

Aside from normal shifts in language due to new slang or technology, adversarial settings for text may encourage some to coin new phrases and use coded language to evade classification

Results

1. Our success rate in fooling perspective rose linearly with the edit distance.
2. Our success rate stayed consistent as the number of queries rose.
3. Our success rate rose slightly as the number of neighbors searched grew.





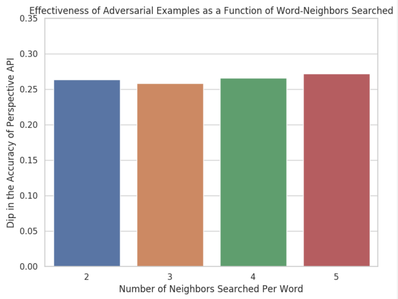


Fig. 2: Effectiveness of our Adversarial Examples

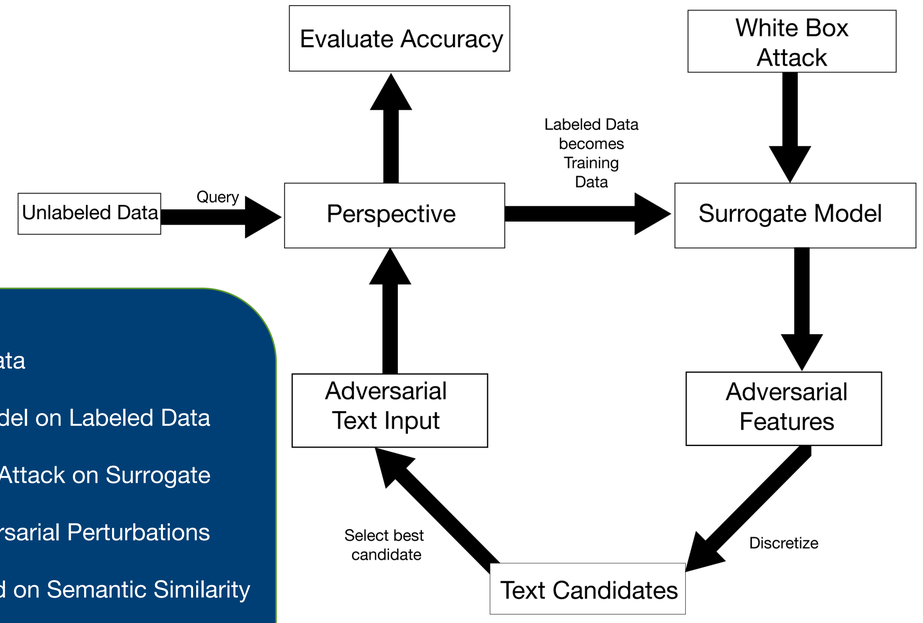
Fig. 3: Text Examples

'"Aww dance now wit your big ass more tears to come when you drop the soap enjoy"	"Aww dance now wit your big but more tears to come when you drop the soap enjoy"
81% Toxic	25% Toxic
"What do you expect Serbia allows their fans to make racist chants at football their men are some of the most ignorant racist people and they see that as acceptable"	"What do you expect Serbia allows their fans to make racist chants at football their men are some of the most not racist people and they see that as acceptable"
76% Toxic	46% Toxic

We successfully fool Perspective and essentially retain the semantic meaning of the comment.

The second comment does flip the classification, however the semantic meaning is murky.

Overview of Pipeline



1. Query Unlabeled Data
2. Train Surrogate Model on Labeled Data
3. Apply a White-Box Attack on Surrogate
4. Discretize the Adversarial Perturbations
5. Choose Input based on Semantic Similarity
6. Evaluate Attack Accuracy

Fig. 4: Pipeline for Generating Adversarial Text Examples

DISCUSSIONS

- The pipeline consistently fails on longer comments because there are too many toxic words.
- Semantic similarity is difficult to retain using just spaCy.
  -- We could use a sent2vec encoder
  -- We could use part of speech tagging

CONCLUSIONS

Online abuse and harassment, though rampant throughout social media platforms can be mitigated through toxicity classification. However, these classifiers can be evaded. Although adversarial machine learning attacks are typically used on continuous data (e.g images), they can be adapted for text. Our pipeline suggests that Perspective, a state of the art classifier, can be fooled by an attacker with no knowledge of the model's internals. Even with only 10,000 queries, changing three words in each comment fools Perspective 25% of the time. However, ensuring that adversarial examples retain semantic similarity requires more work. We hope that by training on these adversarial examples, classifiers can improve their robustness to attacks.

FUTURE WORK

- Modifying pipeline for character level attack
- Testing with different model architectures and parameters
- Place linguistic constraints on candidate words

REFERENCES

[1] . Carlini and D. Wagner. Towards evaluating the robustness of neural networks. arXiv preprint arXiv:1608.04644, 2016.
[2] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. CoRR, abs/1801.04354, 2018.
.[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
[4] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. CoRR, abs/1702.08138, 2017.
[5] . Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pages 506–519. ACM, 2017
[6] S. Samanta and S. Mehta. Towards crafting text adversarial samples. arXiv preprint arXiv:1707.02812, 2017.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation (NSF) Research Experiences for Undergraduates (REU) program. We would like to thank all of the mentors and research fellows at the New York Institute of Technology who have provided their helpful insight and expertise that greatly assisted with our research. We want to extend our thanks to the helpful graduate students, Mahmoud Saleh and Gopi Prasad, for their constant help over the program's duration.

This project is funded by National Science Foundation Grant No. CNS-1559652 and New York Institute of Technology.

www.PosterPresentations.com