We have designed a targeted attack scheme to text classification and rate the “toxicity” of comments. Google’s Perspective seeks to help platforms classify online abuse, trolling and harassment. To tackle this problem, Google launched a project called Perspective which uses machine learning to perform text classification.

Adversarial machine learning attacks to find the optimal perturbation which will fool the model. Since these attacks typically target images, one-pixel perturbations are almost guaranteed to be imperceptible, changing one word can actually obviously not the case for text. Furthermore, whereas rounding. Discretizing some random vector into a 1-word, on the other hand, is more difficult.

The second comment does flip the classification, however the semantic meaning is murky. The pipeline consistently fails on longer comments because there are too many toxic words.

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.

### 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.

### Acknowledgements

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-1559602 and New York Institute of Technology.