

**CHENNAI – PONDICHERRY**

**Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers**

**Abstract**

Although various techniques have been proposed to generate adversarial samples for white-box attacks on text, little attention has been paid to a black-box attack, which is a more realistic scenario. In this paper, we present a novel algorithm, DeepWordBug, to effectively generate small text perturbations in a black-box setting that forces a deep-learning classifier to misclassify a text input. We develop novel scoring strategies to find the most important words to modify such that the deep classifier makes a wrong prediction. Simple character-level transformations are applied to the highest-ranked words in order to minimize the edit distance of the perturbation. We evaluated DeepWordBug on two real-world text datasets: Enron spam emails and IMDB movie reviews. Our experimental results indicate that DeepWordBug can reduce the classification accuracy from 99% to 40% on Enron and from 87% to 26% on IMDB. Our results strongly demonstrate that the generated adversarial sequences from a deep-learning model can similarly evade other deep models.