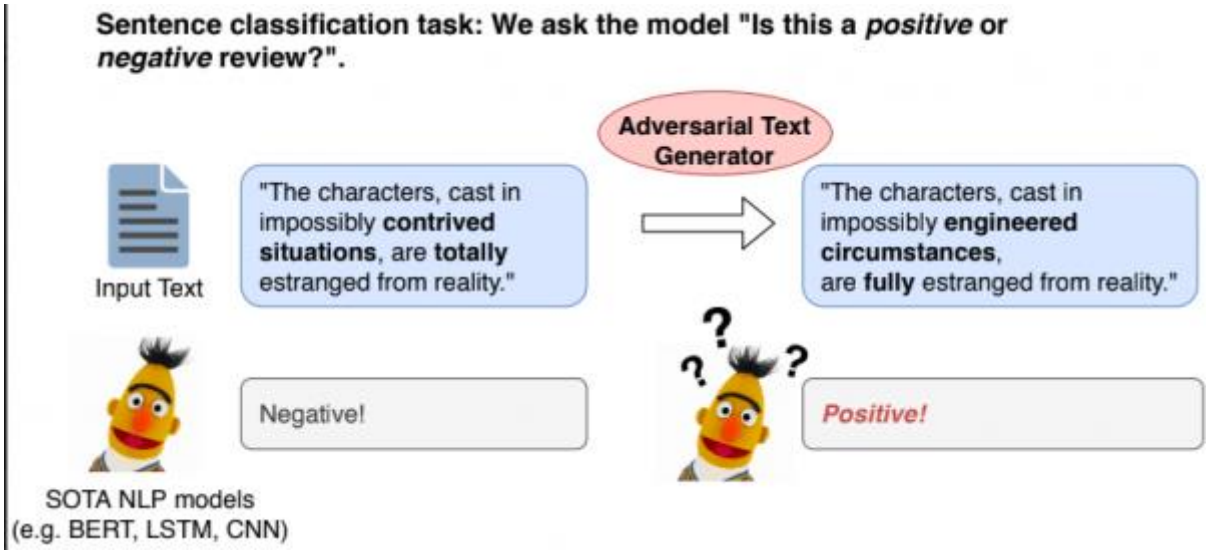
Эксперты разработали систему, способную обмануть модели ИИ от Google



Специалисты Лаборатории компьютерных наук и искусственного интеллекта (CSAIL) Массачусетского технологического института смогли обойти модели искусственного интеллекта, использующие технологии распознавания речи. Разработанная ими система TextFooler способна обманывать модели искусственного интеллекта наподобие Siri и Alexa. В будущем такие системы помогут бороться со спамом и отвечать на нецензурную речь.

TextFooler представляет собой состязательную систему, которые обычно создаются для атак на модели ИИ с использованием технологий распознавания речи с целью выявления недочетов. Система меняет входные предложения путем замены некоторых слов с сохранением смысла и грамматики и с их помощью атакует модель ИИ, чтобы определить, как она обрабатывает измененный текст.

Изменение текста без изменения смысла – задача непростая. Прежде всего TextFooler ищет важные слова, имеющие большую смысловую нагрузку для конкретной модели ИИ, и подбирает для них подходящие синонимы.

По словам специалистов CSAIL, разработанному ими инструменту удалось успешно обмануть три существующие модели, в том числе популярную языковую модель BERT от Google. Изменив лишь 10% текста в предложении, TextFooler смог достигнуть высокого уровня успеха.

Источник: https://www.securitylab.ru/news/504842.php