Contribution ID: **22** Type: **not specified**

Использование методов подсчета неопределенности для борьбы с атаками на детекторы машинно-сгенерированного текста

Saturday, 17 May 2025 15:42 (12 minutes)

В данной работе исследуется применение методов оценки неопределенности для повышения качества детекторов машинно-сгенерированного текста при обработке данных, содержащих атаки, такие как омоглифы, перефразирование и зашумление. Эти атаки не только позволяют обходить детекцию, но и служат для тестирования устойчивости детекторов. Мы проверяем гипотезу о том, что методы оценки неопределенности могут обеспечить более устойчивый подход, устраняя необходимость постоянного дообучения при различных видах атак. Предлагается подход, сочетающий оценку неопределенности с классификаторами на основе скрытых представлений языковых моделей. Эксперименты на датасетах M4GT и RAID демонстрируют конкурентоспособную точность (ROC-AUC 0.8977) при значительно меньших вычислительных затратах по сравнению с тонкой настройкой больших языковых моделей (fine-tuning LLM).

Primary author: LEVANOV, Valeriy (Moscow Institute of Physics and Technology, Moscow)

Co-authors: Ms VOZNYUK, Anastasia (Moscow Institute of Physics and Technology, Moscow); Dr ANDREY, Grabovoy (Moscow Institute of Physics and Technology, Moscow)

Presenter: LEVANOV, Valeriy (Moscow Institute of Physics and Technology, Moscow)

Session Classification: 17-Машинное обучение и нейросети

Track Classification: Машинное обучение и нейросети