УДК 004.912

РАСПОЗНАВАНИЕ СГЕНЕРИРОВАННОГО ТЕКСТА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Скрыльников С.С.
(Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассматривается применение нейронных сетей для классификации сгенерированного текста. Представлено описание архитектур применяемых сетей трансформеров и используемого набора данных.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №621296 «Разработка технологий для персонификации разговорного искусственного интеллекта».

Языковые модели, используемые в настоящее время для генерации текста, достигают отличных результатов во многих видах деятельности, включая изменение стиля текста и даже его генерацию. Однако, этими целями их использование не ограничено, их также применяют для написания фейкового политического контента, новостей и отзывов. В связи с этим актуальной становится задача автоматического распознавания сгенерированного текста.

Для проведения исследований были выбраны датасеты, предоставленные на платформе Kaggle в рамках соревнования «RuATD-2022», они представляют собой корпус размеченных предложений на русском языке. Каждый из них содержит около 215 тысяч примеров, разделённых на тренировочную, тестовую и валидационную выборки. Разница между ними заключается в цели их применения. Первый датасет используется для бинарной классификации и имеет метки H для текстов, написанных человеком и M для сгенерированных предложений. Второй датасет включает в себя 14 классов, так как его целью является мультиклассовая классификация. Каждому сгенерированному тексту в ней соответствует название используемой модели, а текстам, написанным человеком, соответствует метка Human.

Для задач классификации текстов используется большое количество архитектур, самыми отличившимися в этой области на данный момент являются архитектуры моделей трансформеров, такие как Bidirectional Encoder Representations from Transformers (BERT) и Generative Pre-trained Transformers (GPT). Трансформеры направлены на обработку последовательностей данных, в связи с чем становятся часто применимы для работы с текстовой информацией. В данной работе рассматривается вопрос применимости этих моделей для задачи распознавания сгенерированного текста.

В данной работе было реализовано практическое решение задачи распознавания сгенерированного текста на основе моделей-трансформеров BERT и GPT.