УДК 004.89

ИССЛЕДОВАНИЕ СГЕНЕРИРОВАННОГО ТЕКСТА НА ПРЕДМЕТ РАСПОЗНАВАНИЯ ИЗМЕНЕНИЙ СЕРВИСАМИ ИДЕНТИФИКАЦИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Дворников А.С. (ИТМО), **Стрижов Д.А.** (ИТМО), **Унтила А.А.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Федоров Д.А.
(ИТМО)

Введение. Автором разнородных документов или их частей может быть как человек, так и искусственный интеллект (далее - ИИ), например, Large Language Models (LLM). В связи с этим, в процессе анализа определённого документа важны не только его оформление, содержание и логичность, но и его авторство. Это имеет большой вес, если от автора ожидается полностью собственное рассуждение в работе или если оценивается уровень схожести документа, написанного ИИ (далее - AI-документ), с документом, написанным человеком [1,2]. Цель данного исследования - выявить процент изменений в AI-документе, который нужно внести человеку, чтобы система распознавания ИИ не выявила присутствия ИИ в документе.

На сегодняшний день существуют различные интернет-сервисы для проверки документа на уникальность (сервисы антиплагиата), например, сервис antiplagiat.ru [3]. Они работают по принципу поиска заимствований в огромной базе разнородных документов. Большинство таких сервисов не работают с ИИ, они нужны, для того чтобы определить, использовал ли человек фрагменты из чужих текстов для написания своего собственного. Соответственно, основная задача сервисов антиплагиата не состоит в том, чтобы отличать документ, написанный человеком, от AI-документа.

Стоит отметить, что сейчас на сервисе проверки документа на уникальность antiplagiat.ru [3] есть возможность проверки документа на написание ИИ, которая ориентирована в том числе и на анализ работ студентов. Но исходный код данного сервиса закрыт и никаких исследований этого функционала не проводилось.

Помимо сервисов антиплагиата, существуют различные интернет-сервисы, которые заявляют, что способны отличить AI-текст от текста, написанного человеком, например, AI-Text-Classifier [4]. Однако такие системы утверждают об этом в формате вероятности того, что текст написан ИИ, на основе количества ошибок и точности текста.

Основная часть. Стоит отметить, что в одном из исследований было показано, что точность LLM в обнаружении документов, написанных человеком (69-96% в среднем), выше точности обнаружения документов, написанных ИИ (15-77% в среднем). Также было показано, что большинство из существующих сервисов могут почти стопроцентно определить, что текст полностью написан человеком, тогда как если при написании текста использовался ИИ, то такие сервисы не могут точно указать автора. А если в документ, написанный ИИ, внести правки вручную, то процент аккуратности LLM становится ещё меньше. Исходя из этого, не стоит полагаться на точность и надёжность сервисов обнаружения документов, написанных ИИ [1].

В нашем исследовании использовались три ресурса: GPT-2 Output Detector, ZeroGPT и GPT Zero, - так как эти ресурсы доступны, бесплатны и имеют наибольшую аккуратность в проведённом ранее исследовании [1].

В качестве исходных данных были использованы три типа документов:

1. документы, написанные ИИ по определённому запросу;
2. документы, написанные ИИ по определённому запросу и некоторая часть которых (5%-80%) была исправлена/переформулирована человеком вручную;
3. документы, написанные ИИ по определённому запросу и некоторая часть которых (5%-80%) была исправлена/переформулирована ИИ.

В качестве ИИ, который генерировал тексты и по запросу исправлял их, были открытая модель gpt-2 из библиотеки OpenAI, а также бесплатные версии моделей ChatGpt в интернете. Такой выбор был сделан ввиду того, что обычный человек, который хочет проверить документ на присутствие ИИ, в большинстве случаев не будет использовать платные сервисы при наличии бесплатных и более доступных. Таким образом, мы ставим себя на позицию простого человека.

Процент корректировки документа (5-80%) определялся по количеству переформулированных предложений, например, с помощью перестановки слов и замены некоторых слов на синонимы, в отношении к общему количеству предложений в документе.

Всего было сгенерировано 100 документов объёмом около 300 слов с различным содержанием для большей точности исследования. Для каждого запроса ИИ был использован шаблон промпта: «Напиши научную работу на тему <тема> объёмом около 300 слов.» Для проверки каждый из трёх полученных документов проверялся в каждом из трёх сервисов проверки документа на ИИ, а после все результаты записывались в общую таблицу с полученными данными.

Выводы. В итоге исследования был определён процент исправлений, которые нужно сделать человеку в AI-документе, чтобы системы распознавания ИИ не смогли определить тот факт, что ИИ присутствовал при написании документа.

Данный процент позволяет понять, при каком количестве отредактированного текста невозможно отличить документ, написанный человеком, от AI-документа, а также является полезным в разрабатывании алгоритма выявления AI-документов, исправленных человеком, что планируется осуществить в ближайшем будущем.

Список использованных источников:

1. Weber-Wulff D. et al. Testing of detection tools for AI-generated text //arXiv preprint arXiv:2306.15666. – 2023.
2. Dergaa I. et al. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing //Biology of Sport. – 2023. – Т. 40. – №. 2. – С. 615-622.
3. Антиплагиат [Электронный ресурс]. - 2023. - URL: https://docs.antiplagiat.ru/ru/api/ (дата обращения: 22.12.2023).
4. OpenAI Text Classifier [Электронный ресурс]. - 2023. - URL: https://platform.openai.com/ai-text-classifier (дата обращения: 10.12.2023).