УДК 004.89

Метод детекции AI-сгенерированных текстов на основе многопризнакового вектора (стилометрия, семантика, перплексия)

Староверов А.Н., Петров Н.В. (МИЭМ НИУ ВШЭ)
Научный руководитель - кандидат технических наук, профессор Авдошин С.М. (МИЭМ НИУ ВШЭ)

Введение. В настоящее время быстрое развитие и широкий выбор нейросетей затрудняет возможность и людям, и организациям отслеживать, используется ли искусственный интеллект в текстах, с которыми они работают. Однако, существуют области, где это знание необходимо. Например, образование (для проверки работ на списывание), маркетинг(людям хочется знать, не написан ли текст рекламы с помощью ИИ), а также гос.структуры(потому что такие тексты могут распространяться с целью повлиять на граждан и как-либо изменить их мнение). Для борьбы с такими текстами необходим алгоритм, позволяющий определять, сгенерирован текст с помощью ИИ или нет. Существующие модели для детекции часто основаны лишь на одном признаке, что делает их уязвимыми к целенаправленному обходу. Таким образом, возникает необходимость в методе, сочетающим в себе сразу несколько признаков, что сделает его более устойчивым и эффективным в сравнении с аналогами.

Основная часть. Предлагаемый метод основан на построении многопризнакового вектора с последующей классификацией с помощью градиентного бустинга. Решение состоит из нескольких этапов. Во-первых, необходим размеченный набор данных, в котором будет указано, какие тексты являются ИИ-сгенерированными, а какие нет. После того, как данный датасет готов, используется классификатор, который из каждого текста будет извлекать 4 признака: стилометрические – набор из 19 характеристик, которые описывают структуру текста(например, длина предложения, длина слов, число знаков препинания) [1-3], семантические – для анализа связей и глубины текста(используется дообученная на нашем датасете DistilBert[4]), перплексия – вычисляется как разность между перплексией авторегрессионной языковой модели(например, GPT2) и перплексией маскированной языковой модели(например, BERT)[5] и устойчивость текста – для этого текст подвергается переводу, парафразингу с помощью нескольких моделей, после чего вычисляется разность представлений текстов в векторном виде(предполагается, что у ИИ-текстов разность будет достаточно большой из-за того, как устроена генерация таких текстов. После того, как для каждого текста составлен вектор из данных 4 признаков, используется машинное обучение с применением градиентного бустинга (LightGBM[6]), который в итоге находит идеальные веса для каждого признака. Таким образом, когда модель обучена, на вход подается текст, из него выделяются 4 признака, подставляются веса и модель показывает, с какой вероятностью данный текст является ИИ-сгенерированным.

Выводы. В работе предложен новый метод детекции ИИ-сгенерированных текстов, новизна подхода заключается в том, что используется объединение существующих метрик и моделей в рамках одного вектора признаков с последующей оптимизацией весов с помощью градиентного бустинга. Ожидается, что такой метод покажет высокие результаты на тестовой выборке (Accuracy 90%+), а также будет устойчив к различным адаптивным атакам и языковым моделям. Модель, основанную на данном методе, можно будет применять в самых разных задачах: определение фейковых новостей, списывания студентов и школьников, проверка литературы, проверка рекламных текстов, фильтрация постов и сообщений в социальных сетях. Дальнейшие исследования будут направлены на тестирование метода на различных датасетах (включая мультиязычные), а также адаптацию к новым генеративным моделям.

Список использованных источников:

1. Burrows J. ‘Delta’: a measure of stylistic difference and a guide to likely authorship //Literary and linguistic computing. – 2002. – Т. 17. – №. 3. – С. 267-287.
2. Ippolito D. et al. Automatic detection of generated text is easiest when humans are fooled //Proceedings of the 58th annual meeting of the association for computational linguistics. – 2020. – С. 1808-1822.
3. Sanh V. Lysandre Debut, Julien Chaumond, and Thomas Wolf //Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108. – 2019. – Т. 2.
4. Dale R. Natural language generation: The commercial state of the art in 2020 //Natural Language Engineering. – 2020. – Т. 26. – №. 4. – С. 481-487.
5. Xu Z., Sheng V. S. Detecting AI-generated code assignments using perplexity of large language models //Proceedings of the aaai conference on artificial intelligence. – 2024. – Т. 38. – №. 21. – С. 23155-23162.
6. Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree //Advances in neural information processing systems. – 2017. – Т. 30.