Artificial Text Detection

Екатерина Кострыкина и Валерия Лелик

Мотивации

Современные большие языковые модели для генерации текста показывают впечатляющие результаты: они могут составить стихотворение, изменить стиль текстов и даже написать содержательное эссе на свободную тему. Однако такие модели могут использоваться в злонамеренных целях, например для создания фейковых новостей, автоматических обзоров продуктов и фейкового политического контента. Таким образом, возникает новая задача: научиться отличать тексты, написанные человеком, от текстов, сгенерированных генеративными языковыми моделями.

Актуальность

- определение автоматически сгенерированных отзывов (на продукты, товары) -> риск купить некачественный товар
- определение фейковых новостей -> риск дезинформации
- снижение количества информационного мусора

Данные

С Kaggle - https://www.kaggle.com/c/ruatd-2022-bi/data?select=train.csv

train - 129066

val - 21511

test - 64533

Данные

Схема бинарной разметки содержит следующие обозначения:

- H – текст написан человеком
- M – текст сгенерирован автоматически

Схема мультиклассовой разметки содержит следующие обозначения:

- OPUS-MT – текст сгенерирован моделью машинного перевода OPUS
- ruGPT3-Large – текст сгенерирован моделью ruGPT3-Large
- и так далее

Baseline

Код из репозитория соревнования:

https://github.com/dialogue-evaluation/RuATD/blob/main/Baseline.ipynb

- предобученный DeepPavlov ruBERT
- TD-IDF

Метрики оценки

Accuracy - как в бейзлайне

Precision - проверим, насколько надежна модель при классификации Positive-меток

Recall - учтем корректность предсказания всех Positive меток

F1-score - среднее гармоническое значение между precision и recall

План действий

1. Предобработка данных
2. Бинарная классификация (определить, был ли текст сгенерирован автоматически или написан человеком)
3. Мультиклассовая классификация (определить, какая именно модель была использована для генерации данного текста)
4. Оценка качества: если ниже, чем в бейзлайне, пробуем другие архитектуры и векторайзеры

Команда и распределение задач

Екатерина Кострыкина

1. Предобработка данных
2. Бинарная классификация

Валерия Лелик

1. Мультиклассовая классификация
2. Оценка качества

Список литературы

GLTR: Statistical Detection and Visualization of Generated Text - https://arxiv.org/pdf/1906.04043.pdf

Сайт с визуализацией: для каждого слова в тексте высвечивается его предсказуемость: если многие слова слишком предсказуемы, значит, текст сгенерирован искусственно. Для расчета вероятности слова используется GPT-2, для предсказания соседнего слова (и справа и слева) - Bert

Automatic Detection of Machine Generated Text: A Critical Survey - https://arxiv.org/pdf/2011.01314.pdf

Подборка статей, рассказывается о различных подходах к ATD. Можно прочитать про плюсы и минусы каждого подхода и применить их в своем проекте.

Artificial Text Detection via Examining the Topology of Attention Maps - https://arxiv.org/pdf/2109.04825.pdf

Авторы пробуют подход Topological Data Analysis (TDA) для данной NLP задачи. Он является малоизученным в NLP. В применении показал хороший результат. Авторы показали, что простые линейные классификаторы могут давать хороший результат, наравне с BERT.

Trello

https://trello.com/invite/b/TcOuY07F/fcab56554962a94a6a047b9943b64ea5/nn-methods-project