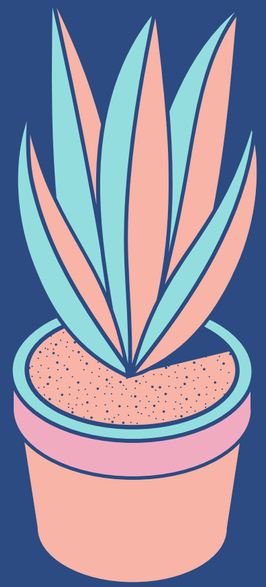
RuATD: распознавание сгенерированных текстов

Катя Волошина & Полина Кудрявцева

Задачи

ПО МОТИВАМ СОРЕВНОВАНИЯ ДИАЛОГ-2022 RUATD

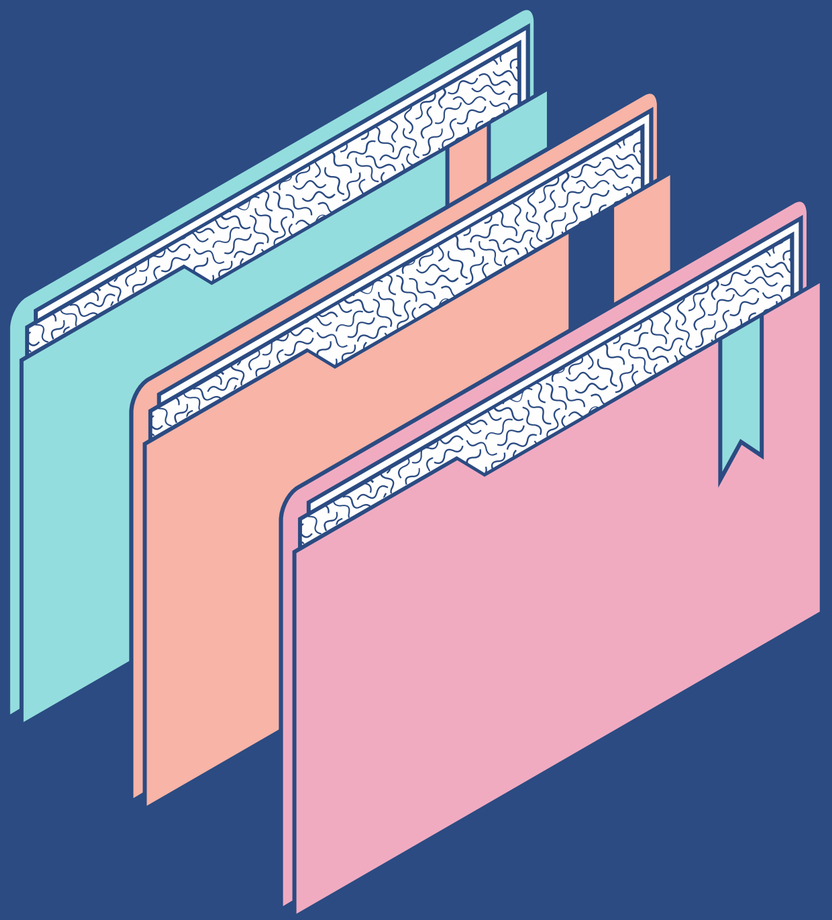


1

Определить, был ли текст сгенерирован автоматически или написан человеком

2

Определить, какая именно модель была использована для генерации данного текста



Мотивация

Модели уже достаточно хорошо генерируют тексты, поэтому могут генерировать фейковые новости, отзывы и др. сообщения в корыстных целях. Важно научиться отличать реальные тексты от фейковых.

Другой задачей будет создать алгоритм, различающий модели. Классификация текстов по моделям, их сгенерировавшим, может помочь оценивать качество генерации текста. В задаче генерации сложно придумать хорошие метрики, поэтому можно считать лучшей моделью ту, которая сложно опознается и часто путается с текстами человека.

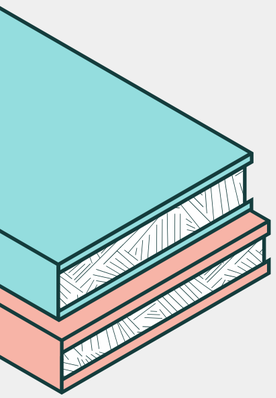
Команда и роли

Катя

- Отвечает за модели
- Прогон бейзлайнов
- Подготовка и тренировка моделей для мультиклассовой классификации

Полина

- SCRUM-мастер (следит за дедлайнами!)
- Препроцессинг данных
- Подготовка и тренировка моделей для бинарной классификации



Данные

- тренировочная выборка: 129065 текстов
- тестовая выборка: 64533 текстов
- валидационная выборка: 21511 текстов
- Тексты, написанные человеком - собраны из открытых источников
- Тексты, сгенерированные моделями
  - M-BART
  - M-BART50
  - M2M-100
  - mT5-Large
  - mT5-Small
  - OPUS-MT
  - ruGPT-3-Large
  - ruGPT3-Medium
  - ruGPT3-Small
  - ruT5-Base
  - ruT5-Base-Multitask
- Обучающая и тестовая выборки размечены автоматически

Бейзлайны

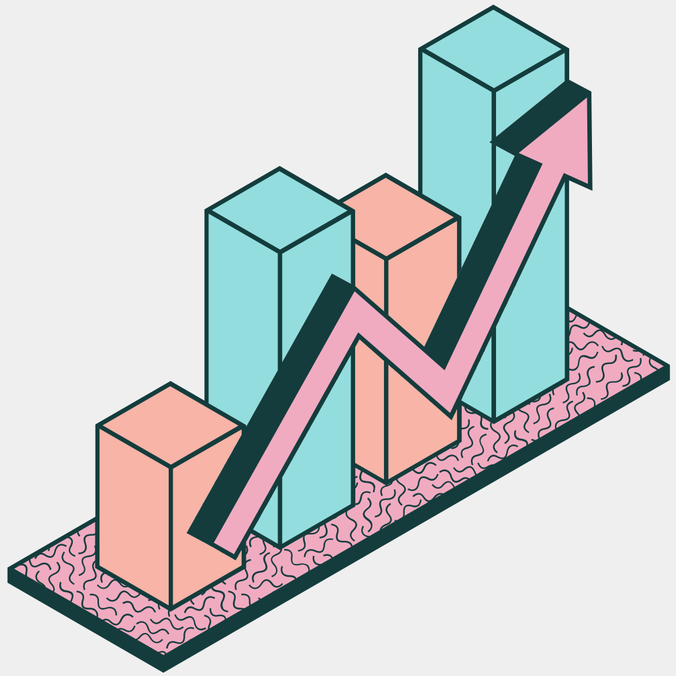
LogReg на TF-IDF

Простейший способ преобразования данных и простейший классификатор

Fine-tuned ruBERT

Это state-off-the-art модель для русского языка. Попробуем дофайнтюнить ее на классификацию текстов.

Метрики



Это задача классификации → воспользуемся классическими метриками.

Accuracy (в мультиклассовой задаче совпадает с micro F1-score): чтобы понимать, сколько правильных ответов дает модель

F1-score (macro): чтобы понимать, насколько сбалансировано модель справляется с задачей (не получается ли высокая accuracy, потому что датасет не сбалансирован и один класс популярней)

План действий

1 ШАГ

Препроцессинг

- токенизация,
- векторизация (Word2Vec, контекстуальные эмбеддинги)
- выделение признаков (длина предложения, частотности би- и триграм, уровень acceptability)

2 ШАГ

Бинарная классификация

- Классификация на машинные тексты и тексты, написанные человеком
- MLP
- BiLSTM
- BERT

3 ШАГ

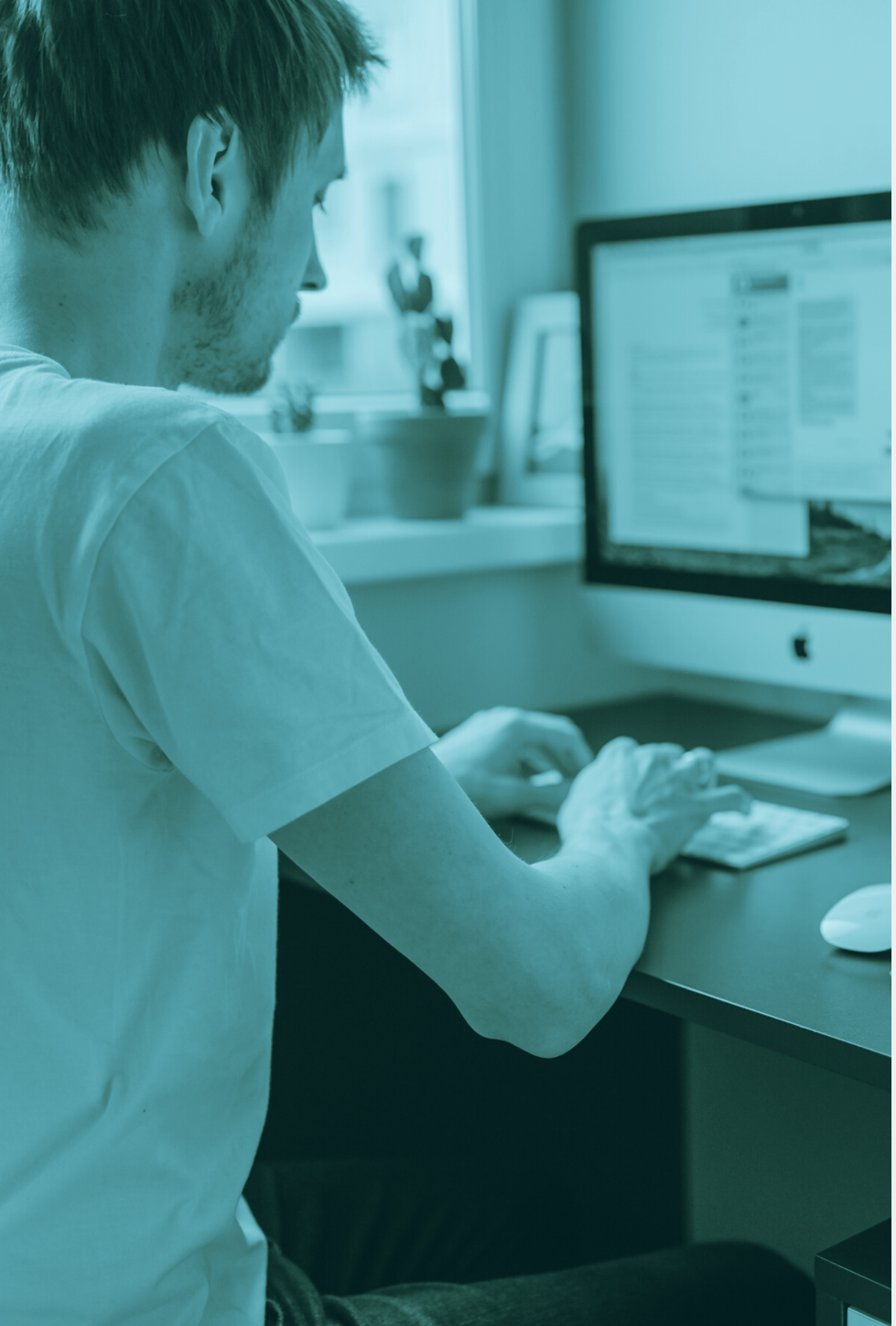
Мультиклассовая классификация

- Классификация машинных текстов по источнику
- Попробуем те же модели, что и для бинарной
- Используем маски, чтобы считать правильность предложения

4 ШАГ

Оценка

- Оценка моделей на тестовой выборке
- Визуализация результатов
- Анализ ошибок моделей



Список литературы

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020, April). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In International Conference on Advanced Information Networking and Applications (pp. 1341-1354). Springer, Cham.
Авторы использовали Grover, GTLP и OpenAI GPT-2 detector для обнаружения сгенерированных текстов. Мы планируем посмотреть, что из этого можно и стоит использовать в нашем случае.

Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M. A., & Szlam, A. (2019). Real or fake? learning to discriminate machine from human generated text. arXiv preprint arXiv:1906.03351.
Не подошло! Мы не хотим использовать генеративные модели в нашем решении.

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. Plos one, 16(5), e0251415.
В статье предложены 13 способов обнаружения сгенерированных текстов (в том числе, наш бейзлайн).

Kushnareva, L., Cherniavskii, D., Mikhailov, V., Artemova, E., Barannikov, S., Bernstein, A., ... & Burnaev, E. (2021). Artificial Text Detection via Examining the Topology of Attention Maps. arXiv preprint arXiv:2109.04825.
Взяли ссылки на другие исследования

Lau, J. H., Armendariz, C., Lappin, S., Purver, M., & Shu, C. (2020). How furiously can colorless green ideas sleep? sentence acceptability in context. Transactions of the Association for Computational Linguistics, 8, 296-310.
Взяли то, как считать acceptability score

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2019). Masked language model scoring.