Соревнование RuATD-2022

https://www.dialog-21.ru/evaluation/2022/ruatd/

Матяш Дарья
Кошкина Ксения

Постановка задачи

Обучить модель отличать тексты, написанные человеком, от сгенерированных (задача бинарной классификации).

Описание данных

Пример разметки

Обозначения двух классов -

текст написан:

- H – человеком
- M – машиной

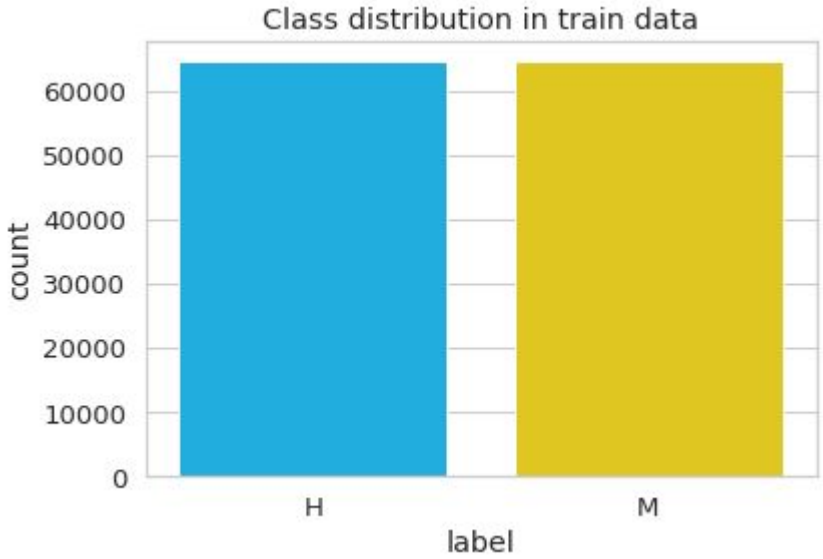
H	M-MT (FR→RU)
Эх, у меня может быть и нет денег, но у меня всё ещё есть гордость.	Может, у меня нет денег, но у меня всегда есть гордость.
Меня покусали комары.	Меня похитили муски.
Я не могу чувствовать себя в гостинице как дома.	Я не могу чувствовать себя дома в отеле.
Эта книга показалась мне интересной.	Я нашёл эту интересную книгу.
Я был полон решимости помочь ему, даже рискуя собственной жизнью.	Я был готов помочь ему в опасности своей жизни.
Моя квартира находится меньше чем в пяти минутах пешком от станции.	Моя квартира находится на расстоянии менее пяти минут от станции.

Данные

Train data - 129066 текстов

Test data - 64533 текста

Validation data - 21511 текст



Для дальнейших экспериментов использовали только 60% train и validation data. Соотношение двух классов в датасетах было сохранено.

Метрики

организаторы ориентируются на **accuracy**

+ смотрели на **precision, recall, F1-score** для более глубокого понимания работы модели

Команда, роли

Даша, Ксюша:

- анализ данных
- генерация идей (в feature engineering, поиске полезной информации)

Ксюша:

- ML-эксперименты

Даша:

- DL-эксперименты

Baseline

	name	metric_acc
0	Baseline Bert	0.79622
1	Baseline Tf-Idf	0.63562

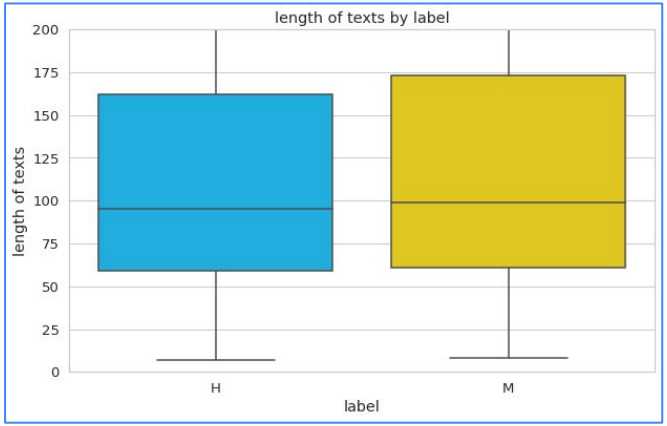
Интересные наблюдения

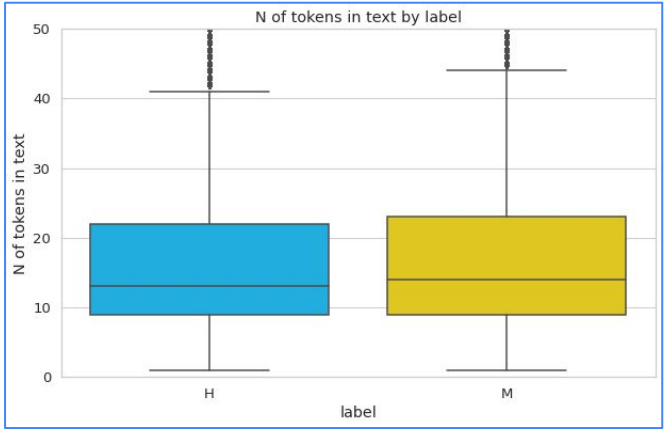
длина текста

искусственно созданный текст чаще длиннее человеческого

число токенов в текстах

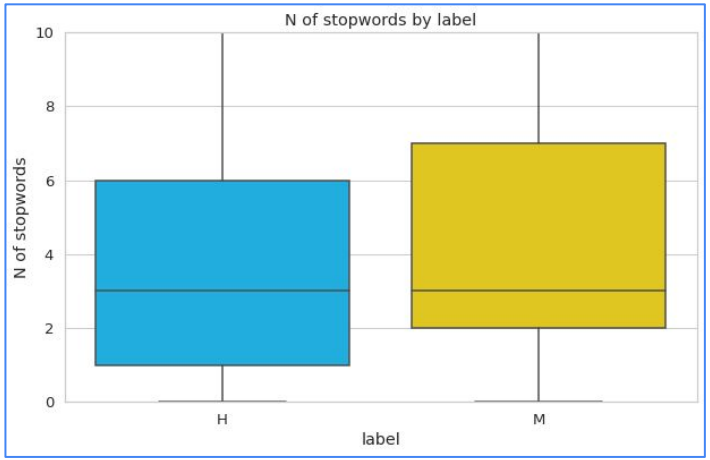
в искусственных текстах скорее больше слов, чем в человеческих



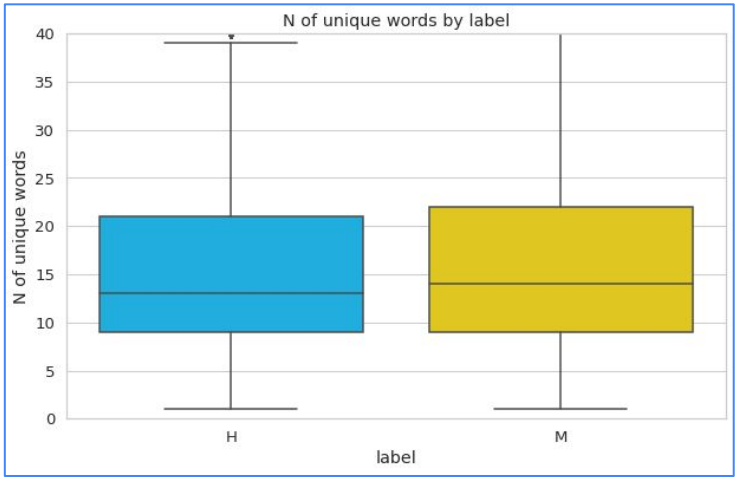


Интересные наблюдения

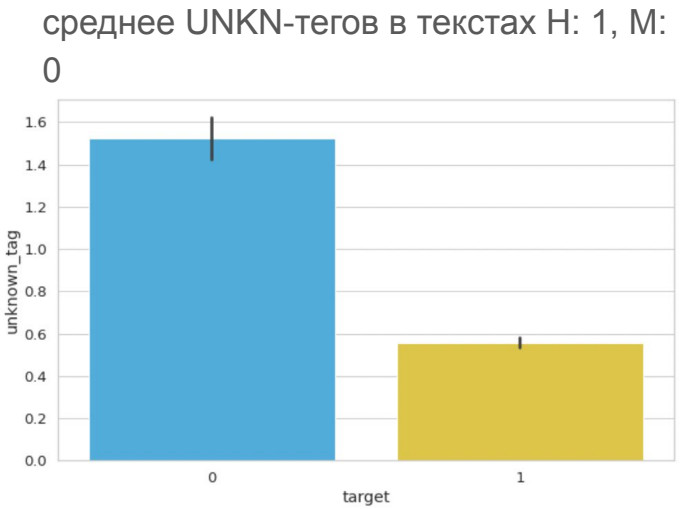
число стоп-слов



число уникальных слов



Интересные наблюдения



<u>В искусственном тексте больше:</u>

- стоп-слов
- уникальных слов
- слов в целом
- длина всего текста
- прилагательных (и меньше глаголов)
- слов, которым pymorphy не может определить POS-тег

PS.:многие из этих наблюдений привели к значительному улучшению качества (см. далее)

Эксперименты

Обучено моделей: больше 10

Общее время обучения моделей: больше 50 часов

Машинное обучение

Улучшенная логистическая регрессия (добавлены фичи):

	precision	recall	f1-score	support
0	0.67	0.69	0.68	6453
1	0.68	0.66	0.67	6453
accuracy			0.68	12906
macro avg	0.68	0.68	0.68	12906
weighted avg	0.68	0.68	0.68	12906

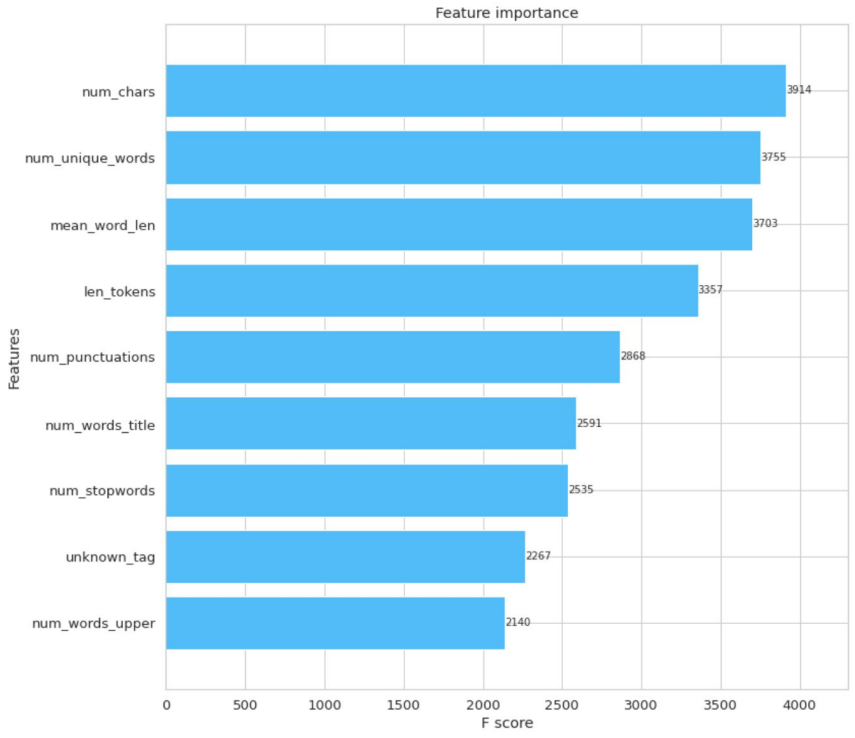
1) Количество None пос-тегов

2) Количество токенов в тексте

3) Средняя длина токена в тексте

Качество: + 0.05 по сравнению с бейзлайном

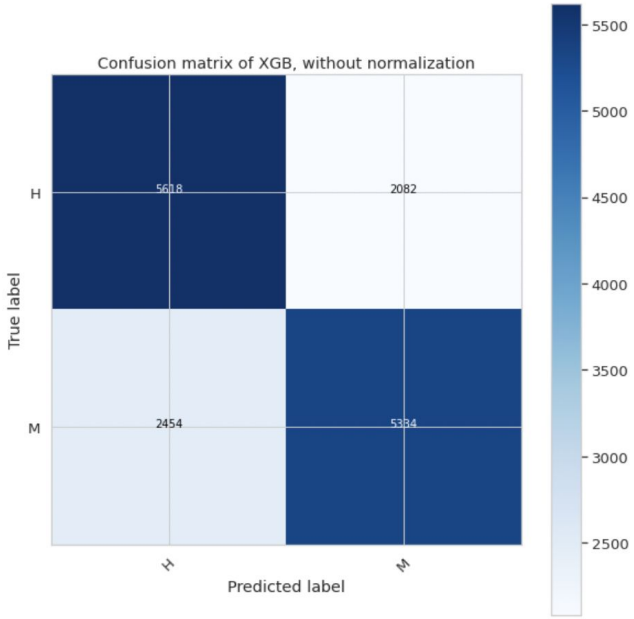
XGBoost



При обучении XGBoost были добавлены дополнительные фичи:
количество символов, количество уникальных слов, средняя длина слова, количество токенов, количество пунктуационных знаков, количество слов с заглавной буквы, количество стоп-слов, количество unknown pos-tags, количество слов в upper case

Первая модель была обучена только на них

XGBoost (добавлены фичи):



1) Количество None пос-тегов

2) Количество токенов в тексте

3) Средняя длина токена в тексте

Качество: + 0.08 по сравнению с

```
Classification Report
              precision    recall  f1-score

           H       0.70      0.73      0.71
           M       0.72      0.68      0.70

    accuracy                           0.71
   macro avg       0.71      0.71      0.71
weighted avg       0.71      0.71      0.71
```

Также были обучены, но безрезультатно:

1) PassiveAgressive Classifier

	precision	recall	f1-score
H	0.54	0.50	0.52
M	0.53	0.57	0.55
accuracy			0.53
macro avg	0.53	0.53	0.53
weighted avg	0.53	0.53	0.53

2) MultinominalNB

	precision	recall	f1-score
H	0.59	0.80	0.68
M	0.69	0.46	0.55
accuracy			0.63
macro avg	0.64	0.63	0.61
weighted avg	0.64	0.63	0.61

BERT family

Ниже представлены топ-4 самых лучших результатов экспериментов

Мы экспериментировали с:

- конкатенацией эмбеддинга [CLS] токена с последнего слоя
- weight_decay
- scheduler
- разные параметры при tokenizer и др.

model name	num_epochs	Accuracy
RuBERT-large	3	0.793
RuBERT-tiny	5	0.7934
multilingual-bert-base-uncased + CLS	3	0.81
XLM-RoBERTa	3	**0.814**

Результат

NB.: использовали только 60% train и validation данных, тем не менее, получилось так:

Baseline	Лучшая модель	Accuracy	Accuracy относительно Baseline
Tf-idf + LogReg	XGBoost	0.71	**+0.08**
BERT	XLM-RoBERTa	0.81	**+0.02**

Анализ результатов

Из ML-части:

- Кастомные фичи num_unique_words, len_tokens, num_chars имели наибольший вес при обучении

И DL-части интересно было:

- частое пересечение предсказаний для ruBERT (tiny и “стандартного”)
- детекция более синтаксически “некрасивых” текстов как H (человек) всеми моделями

Предложения об улучшениях

- большее количество эпох
- исходное количество данных
- подбор гиперпараметров при обучении
- дополнительные классификаторы поверх выхода другого
- ансамбли моделей

Список литературы

1. Santiago Alonso-Bartolome, Isabel Segura-Bedmar. Multimodal Fake News Detection. arXiv:2112.04831 [cs.CL] 9 Dec 2021
2. Bimal Bhattarai, Ole-Christoffer Granmo, Lei Jiao. Explainable Tsetlin Machine framework for fake news detection with credibility score assessment. arXiv:2105.09114v1 [cs.CL] 19 May 2021
3. Sushma Kumari. NoFake at CheckThat! 2021: Fake News Detection Using BERT. arXiv:2108.05419 [cs.CL] 11 Aug 2021
4. Mateusz Szczepański1 Marek Pawlicki Rafał Kozik & Michał Choraś. New explainability method for BERT-based model in fake news detection. 8 Dec 2021
5. Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, Jian Yin. Neural Deepfake Detection with Factual Structure of Text. 15 Oct 2020
6. Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, Maurizio TesconiTweepFake: about Detecting Deepfake Tweets. 6 May 2021

Ссылка на гитхаб

Проект