

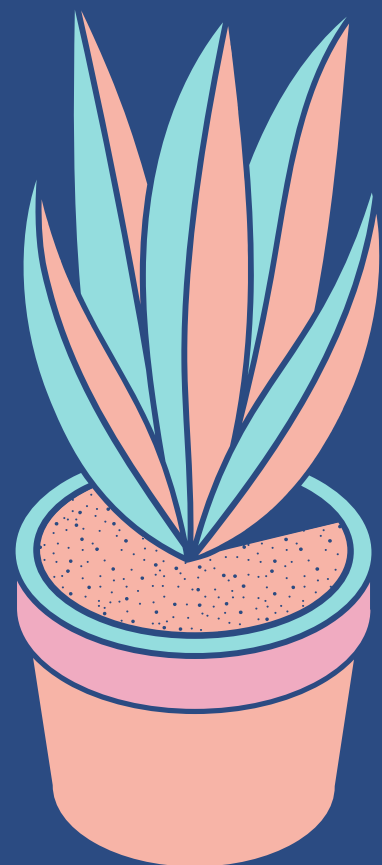


RuATD: распознавание сгенерированных текстов

Катя Волошина & Полина Кудрявцева

Задачи

ПО МОТИВАМ
СОРЕВНОВАНИЯ
ДИАЛОГ-2022 RUATD



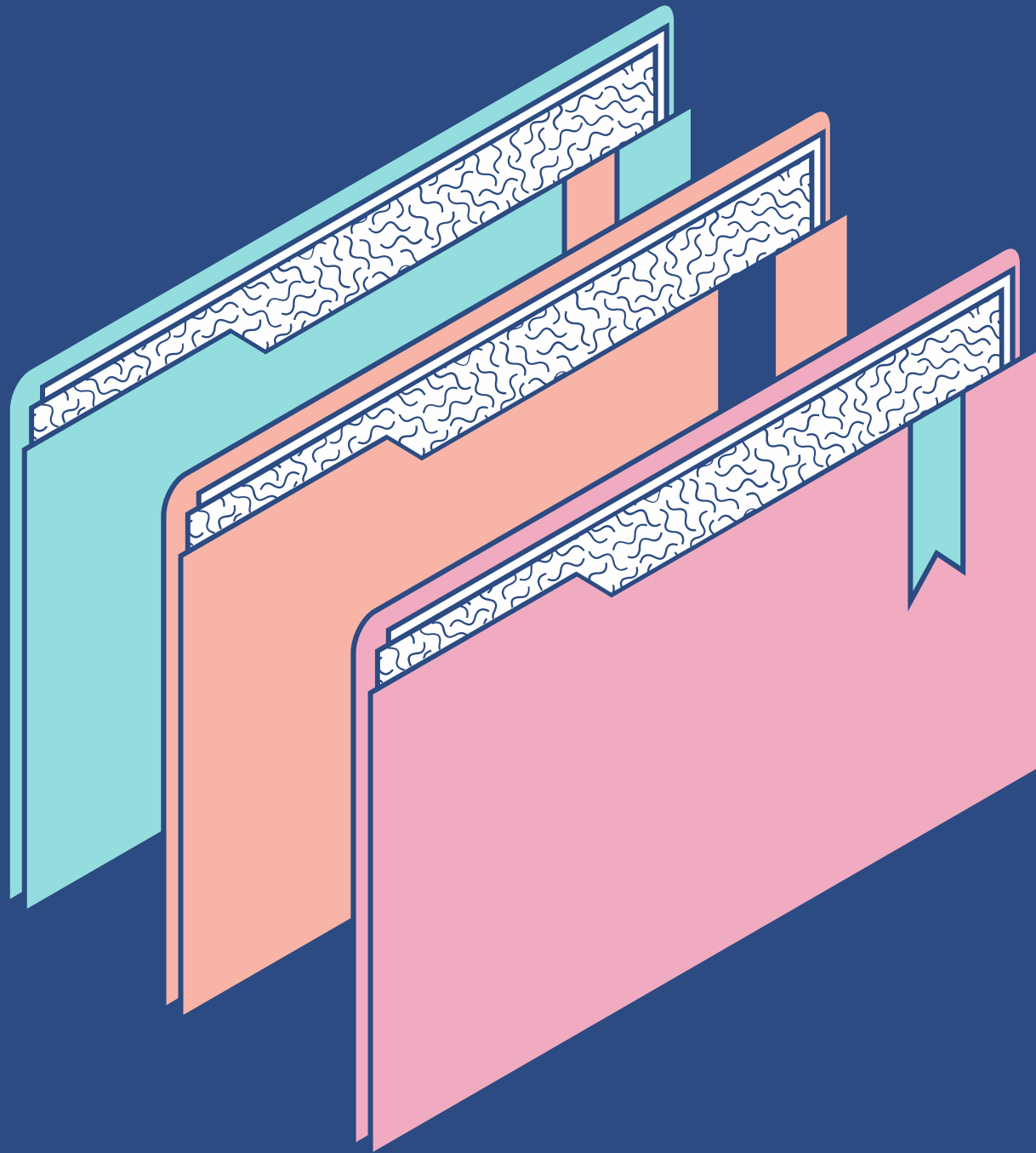
1

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

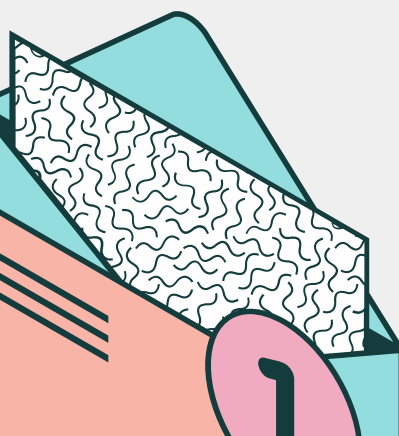
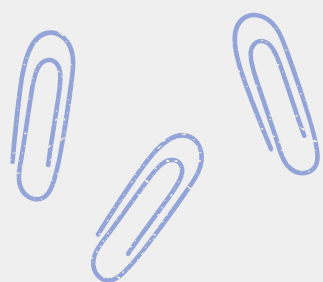
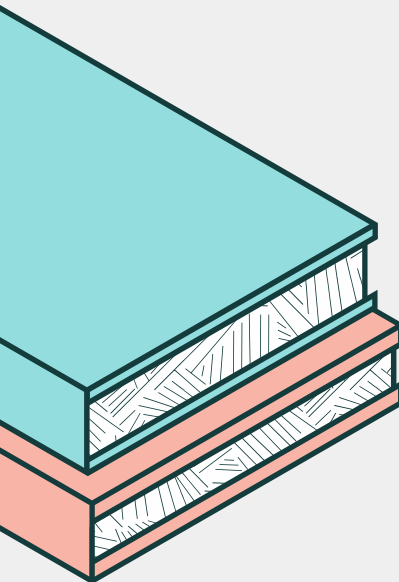
- подбор моделей для обучения
- подготовка и тренировка моделей для мультиклассовой классификации
- подготовка и тренировка моделей для бинарной классификации

Полина

- SCRUM-мастер (следит за дедлайнами!)
- подбор идей для препроцессинга данных и препроцессинг данных
- построение графиков для анализа результатов

Данные

- тренировочная выборка: 129065 текстов
- валидационная выборка: 21511 текстов
- Тексты, написанные человеком - собраны из открытых источников (Википедия, соц. сети и др.)
- 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 авторами соревнования

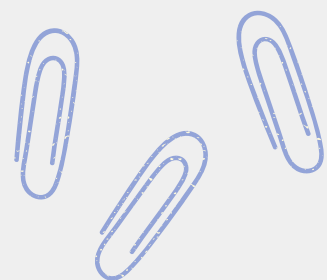
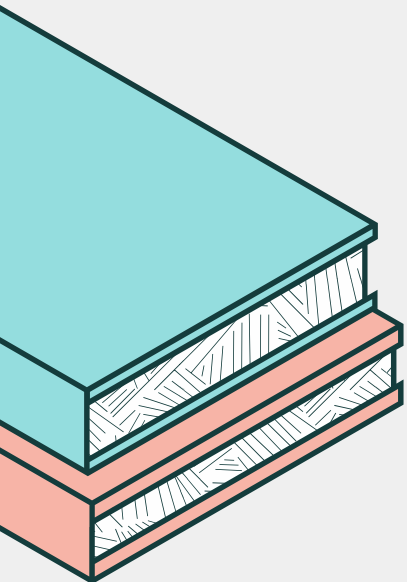


Фрагмент данных

binary разметка

multi разметка

Эх, у меня может быть и нет денег, но у меня всё ещё есть гордость.	Н	Н
Меня покусали комары.	Н	Н
Меня похитили муски.	М	Opus-MT
Я был готов помочь ему в опасности своей жизни.	М	Opus-MT
Моя квартира находится меньше чем в пяти минутах пешком от станции.	Н	Н

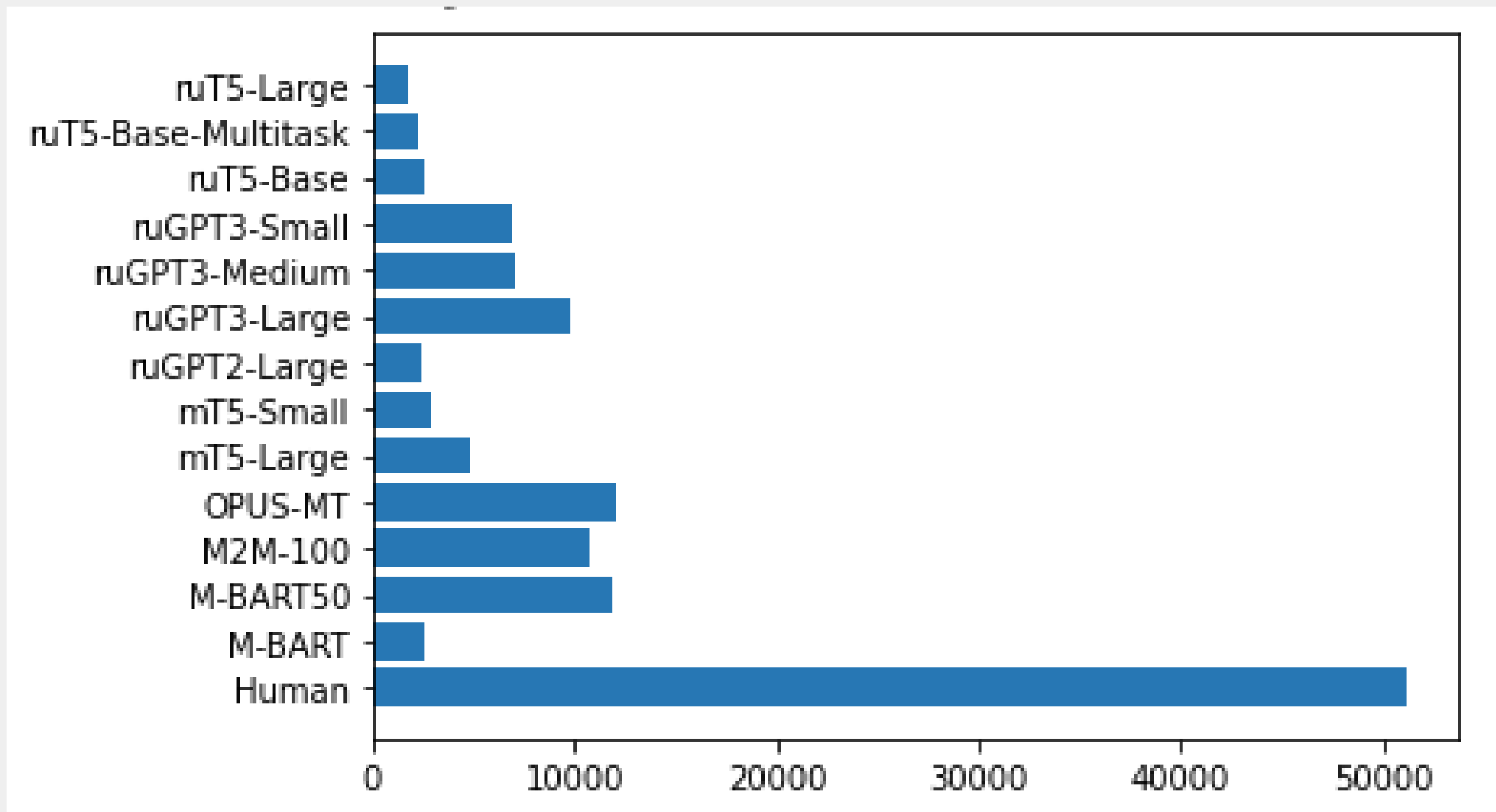


Данные

- выборка binary сбалансированная: примерно 50/50 текстов, написанных человеком, и текстов, сгенерированных моделью
- например, в train binary данных:
 - human: 10756 текстов
 - machine: 10755 текстов
- по длинам предложения распределены так:

	human	machine
средняя длина в токенах	30,07	31,85
средняя длина в символах	221,47	236,86

- binary данные сбалансированы, а multi классы сбалансированы только в отношении машинных и человеческих текстов, но внутри машинных текстов баланс по классам не соблюдается



Бейзлайны


LogReg на TF-IDF

Простейший способ преобразования данных и простейший классификатор

Fine-tuned ruBERT

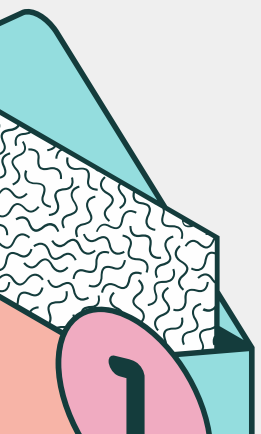
Это state-of-the-art модель для русского языка. Попробуем дофайнтюнить ее на классификацию текстов.



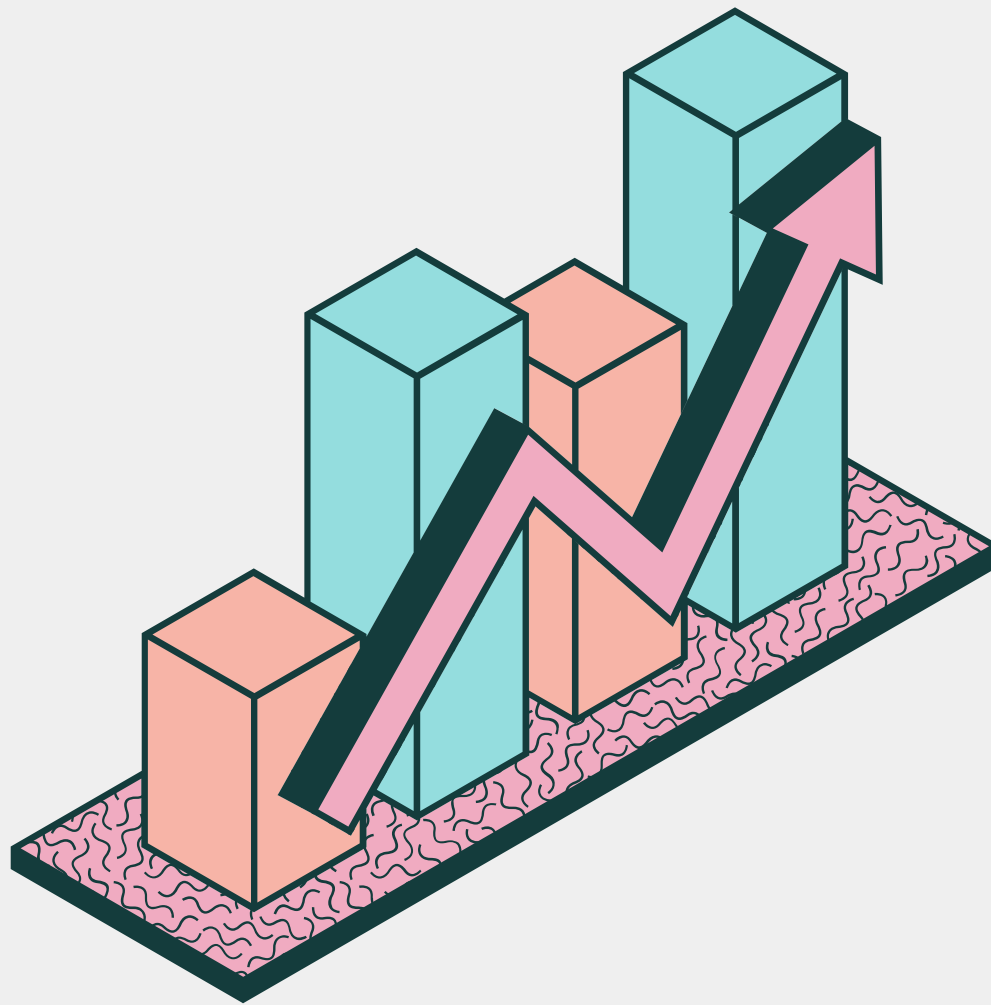


Качество бейзлайнов на мультиклассовой классификации

ruBERT	0.59914
TF-DF + LogReg	0.44300



Метрики



Это задача классификации → воспользуемся классическими метриками.

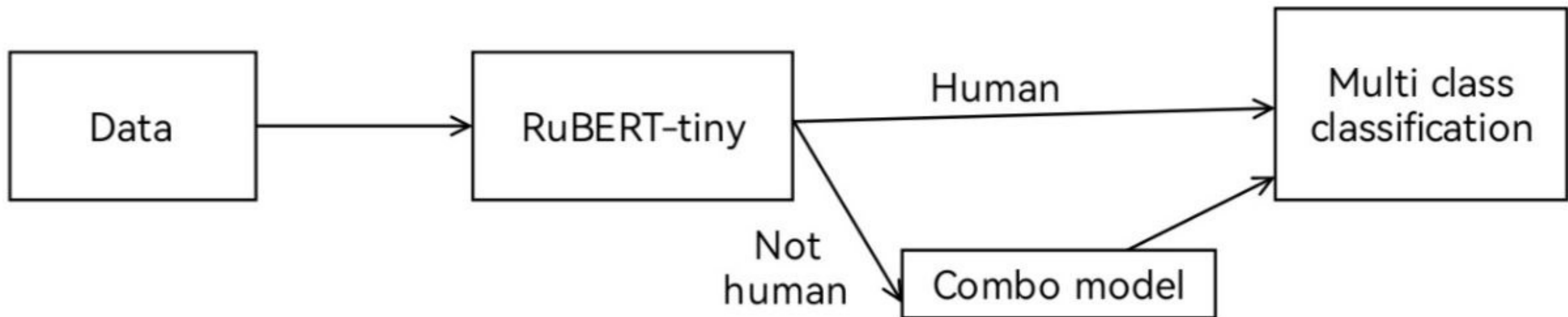
Accuracy (в мультиклассовой задаче совпадает с micro F1-score): чтобы понимать, какова доля правильных ответов у модели

F1-score (macro) для мультиклассовой классификации: посмотреть, как дисбаланс классов влияет на качество модели)

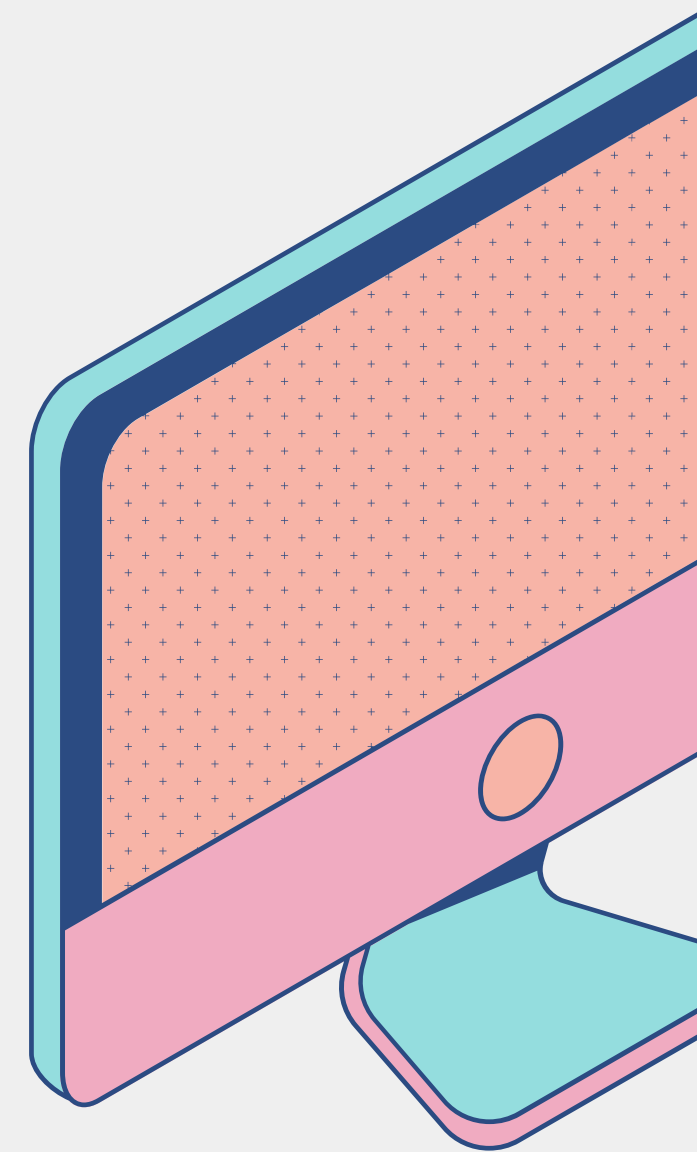
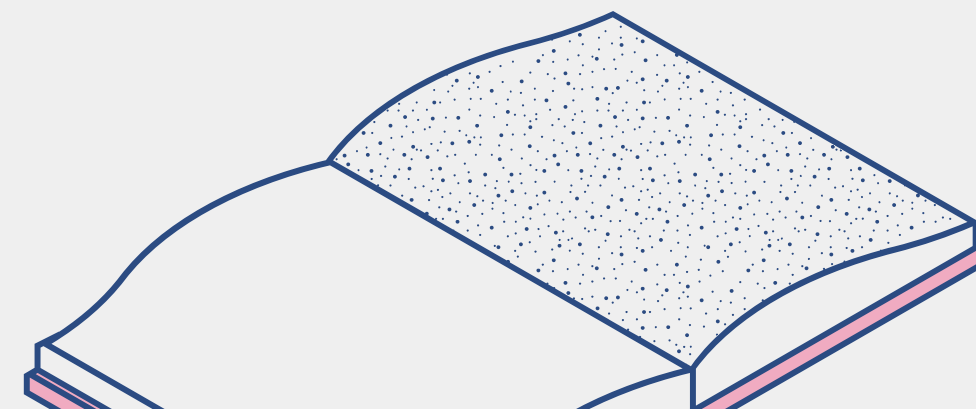
Подготовка данных

- Токенизация, `py morphology2`
- Лемматизация, `py morphology2`
- Векторизация: TF-IDF, `word2vec`
- Выделение признаков у предложений:
 - длина в символах и в токенах
 - кол-во гласных на символы
 - кол-во пробелов на символы
 - кол-во пунктуационных знаков на символы
 - средняя длина слов
 - кол-во длинных слов (≥ 10 букв) на все слова
 - кол-во коротких слов (≤ 3 букв) на все слова
- readability:
 - Dale-Chall
 - Gunning-Fog
 - Flesch
- морфология:
 - кол-во служебных частей речи (ч.р.) на все слова
 - кол-во самостоятельных ч.р. на все слова
 - кол-во именных ч.р. на все слова
 - кол-во глагольных ч.р. на все слова
 - отношение кол-ва именных ч.р. к глагольным

Двухэтапный пайплайн

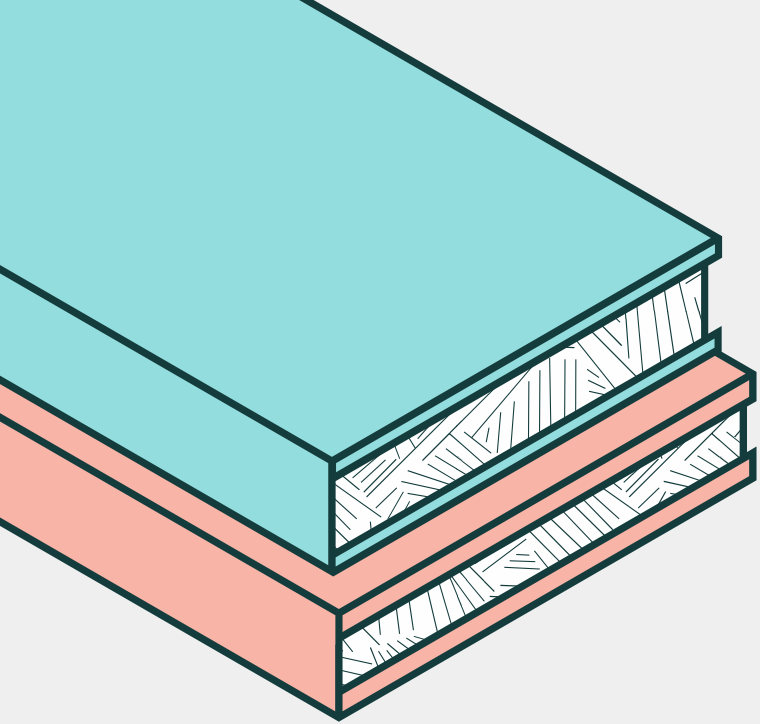


Почему мы выбрали RuBERT-tiny?

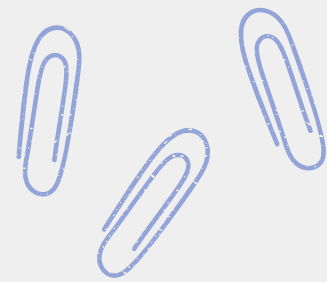


Результаты

encoding	classifier	binary: accuracy
BoW+TF-IDF	FNN	0.73
word2vec	LSTM	0.59
character	CNN	0,66
Native LM	fine-tuning	0.81
text features	FNN	0.6
text features	Logreg	0.61
text features	RandomForest	0.64



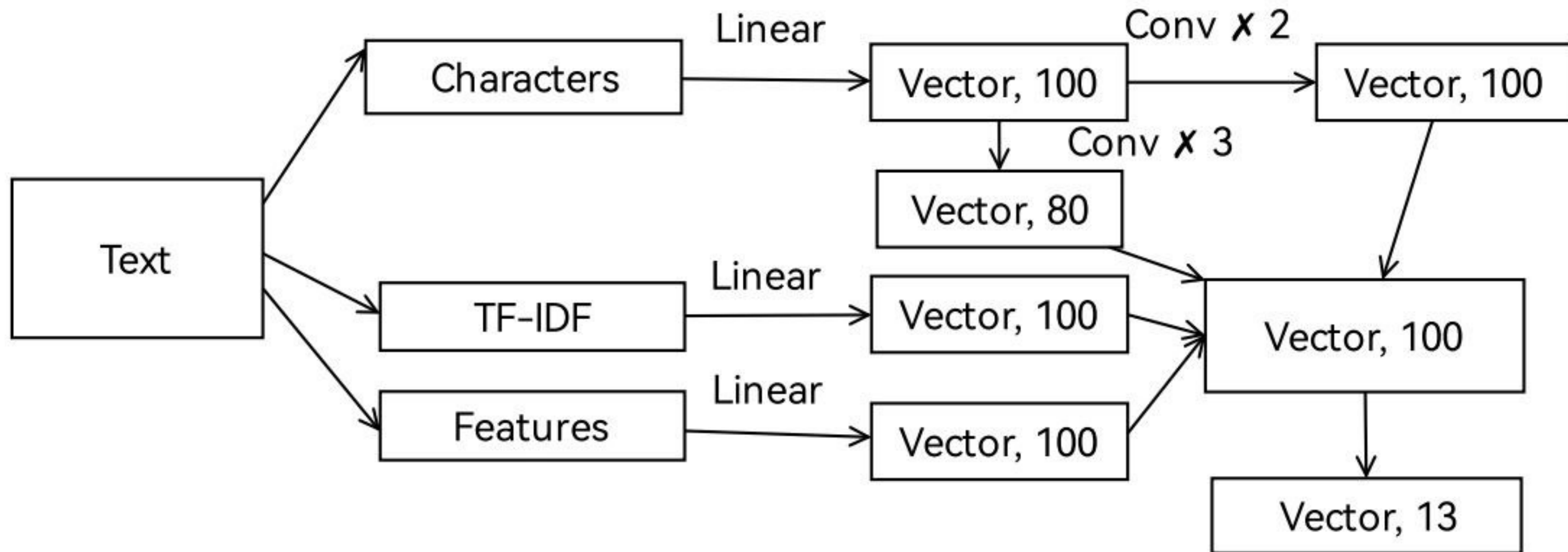
Почему мы выбрали
Combo model?
И что это вообще?



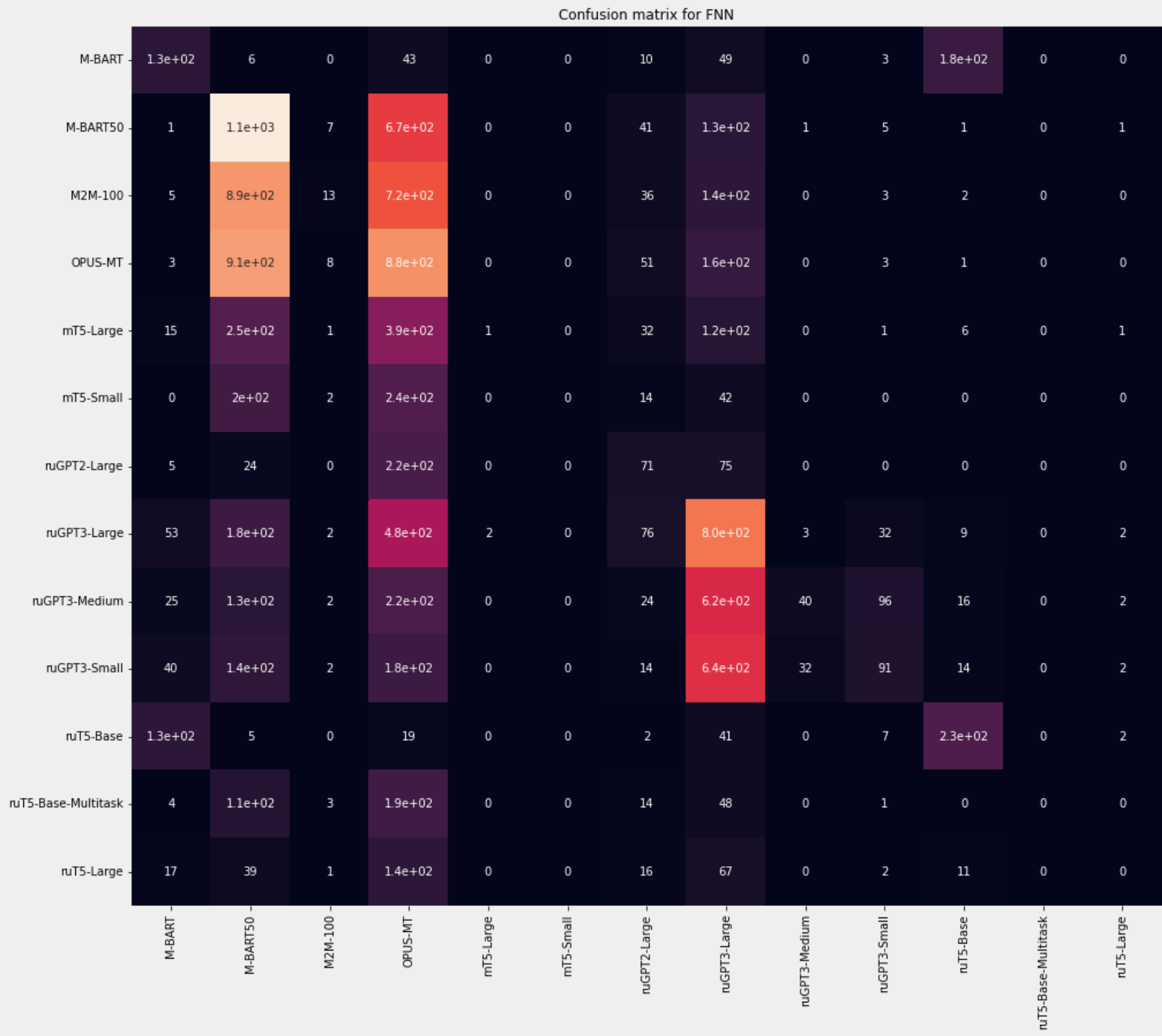
Результаты

encoding	classifier	accuracy	macro f1-score
TF-IDF	FNN	0.3	0.29
word2vec	LSTM	0.29	0.25
character	CNN	0.37	0.32
text features	FNN	0.26	0.17
text features	Logreg	0.22	0.1
text features	RandomForest	0.27	0.24
Combo Model		0.39	0.35

Combo model



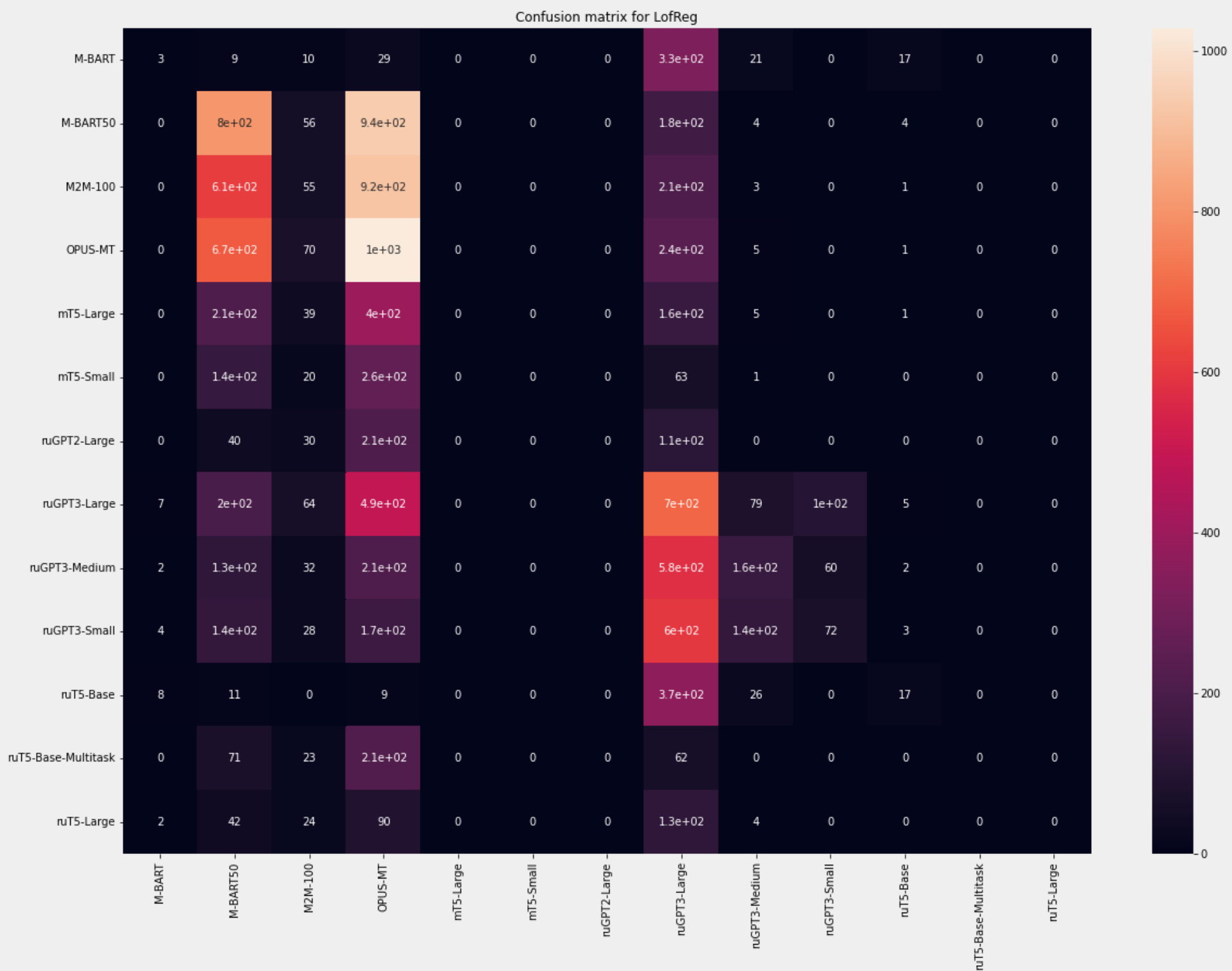
FNN



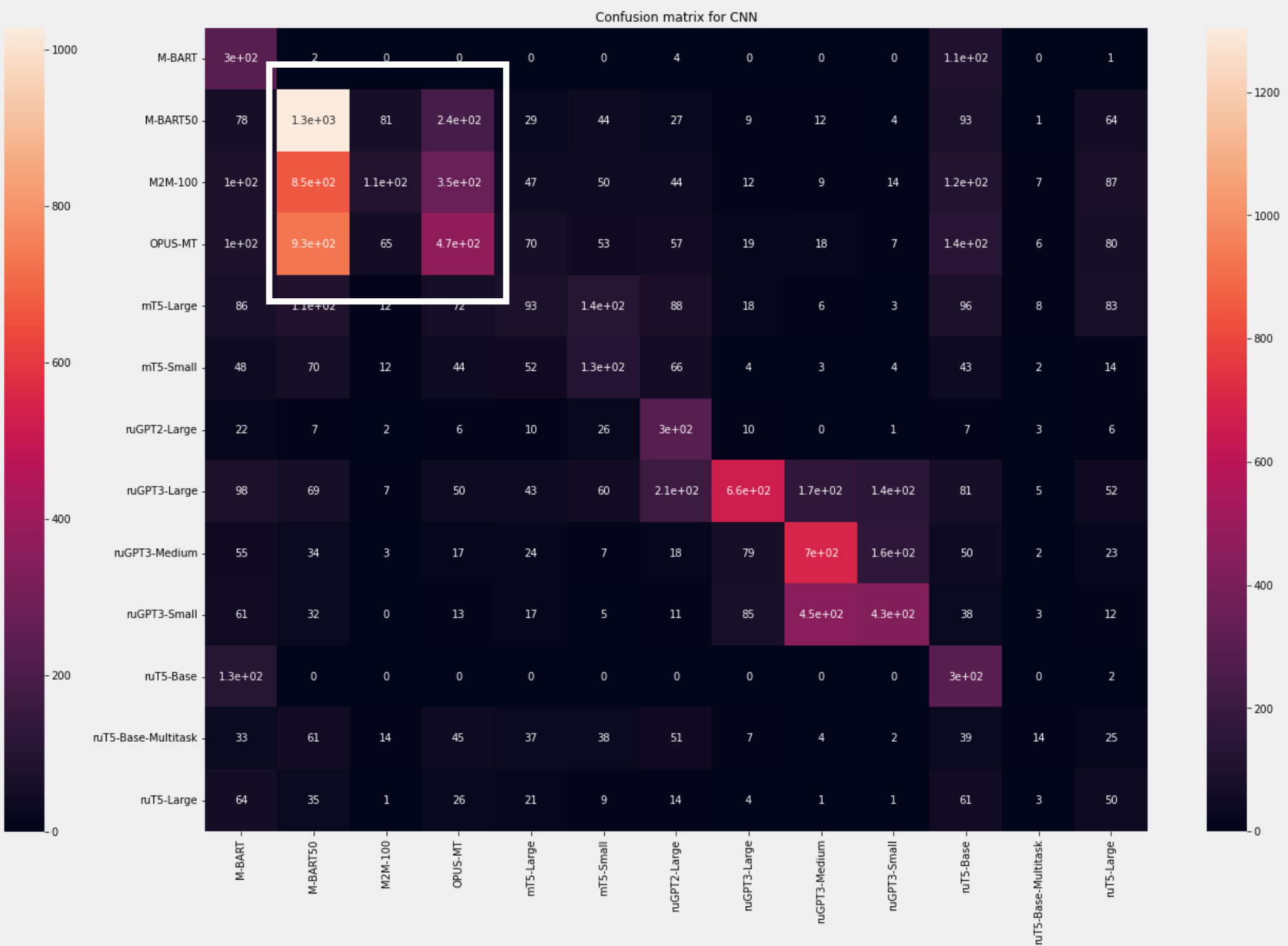
Random Forest



LogReg



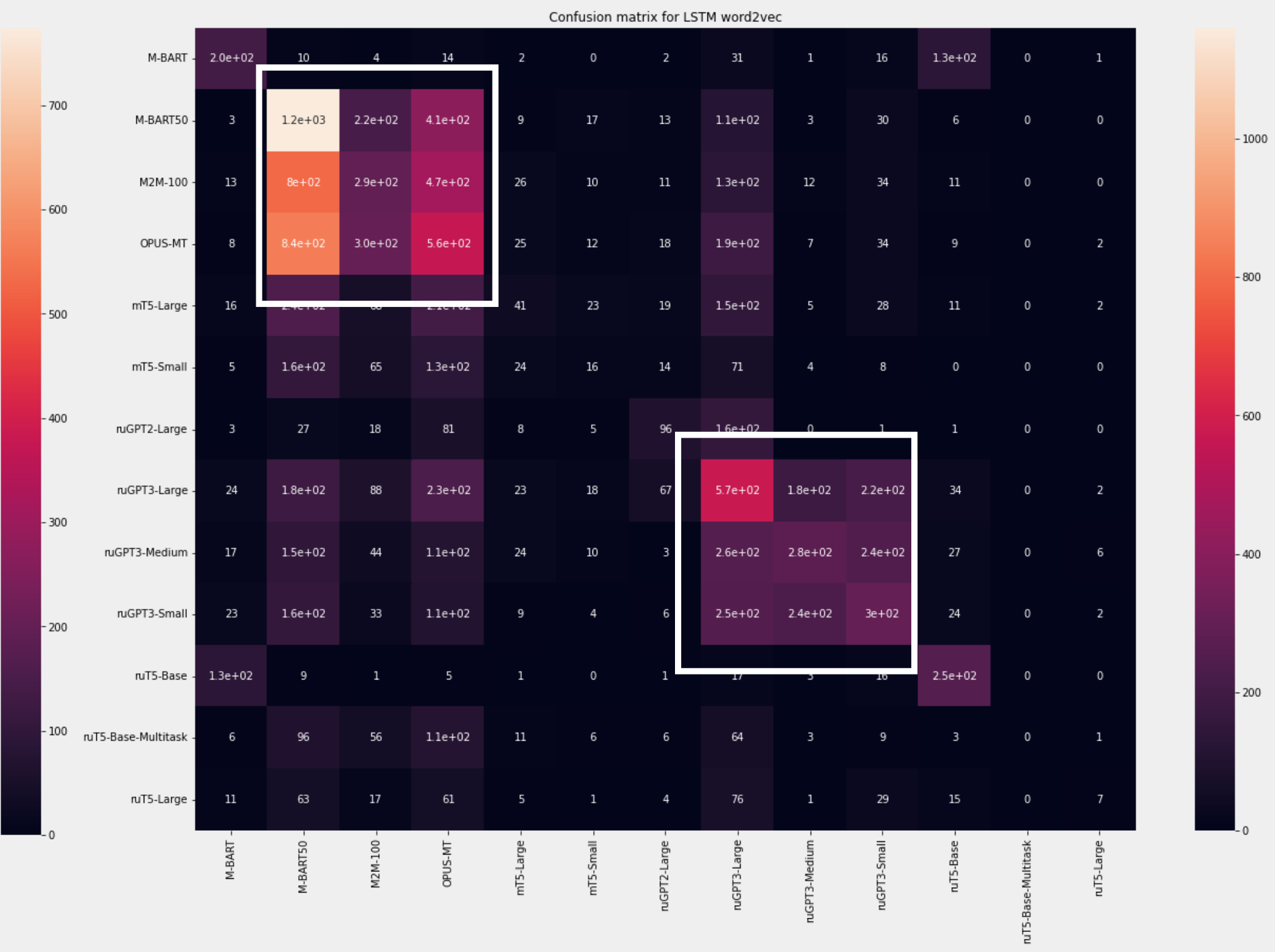
CNN



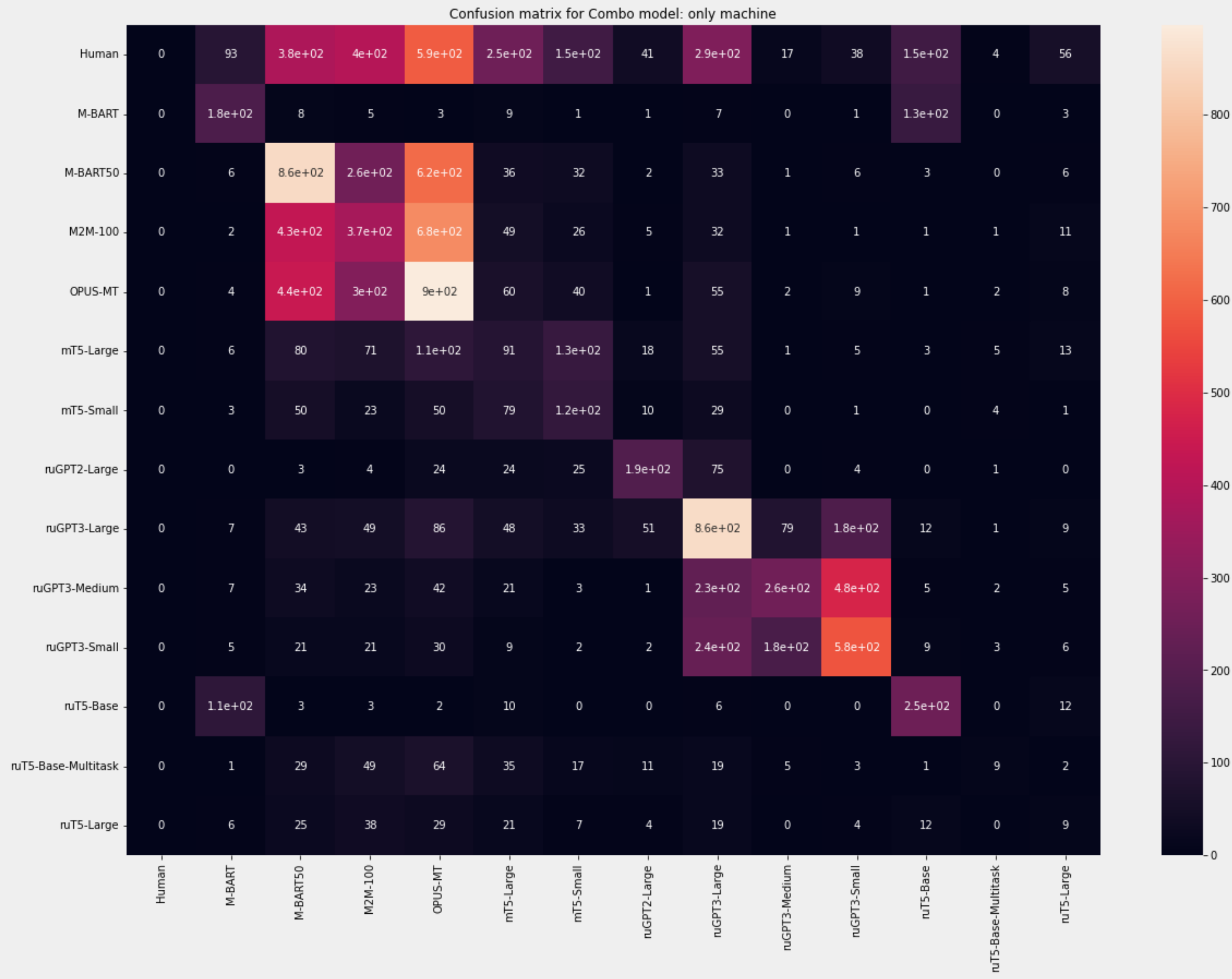
TF-IDF



LSTM



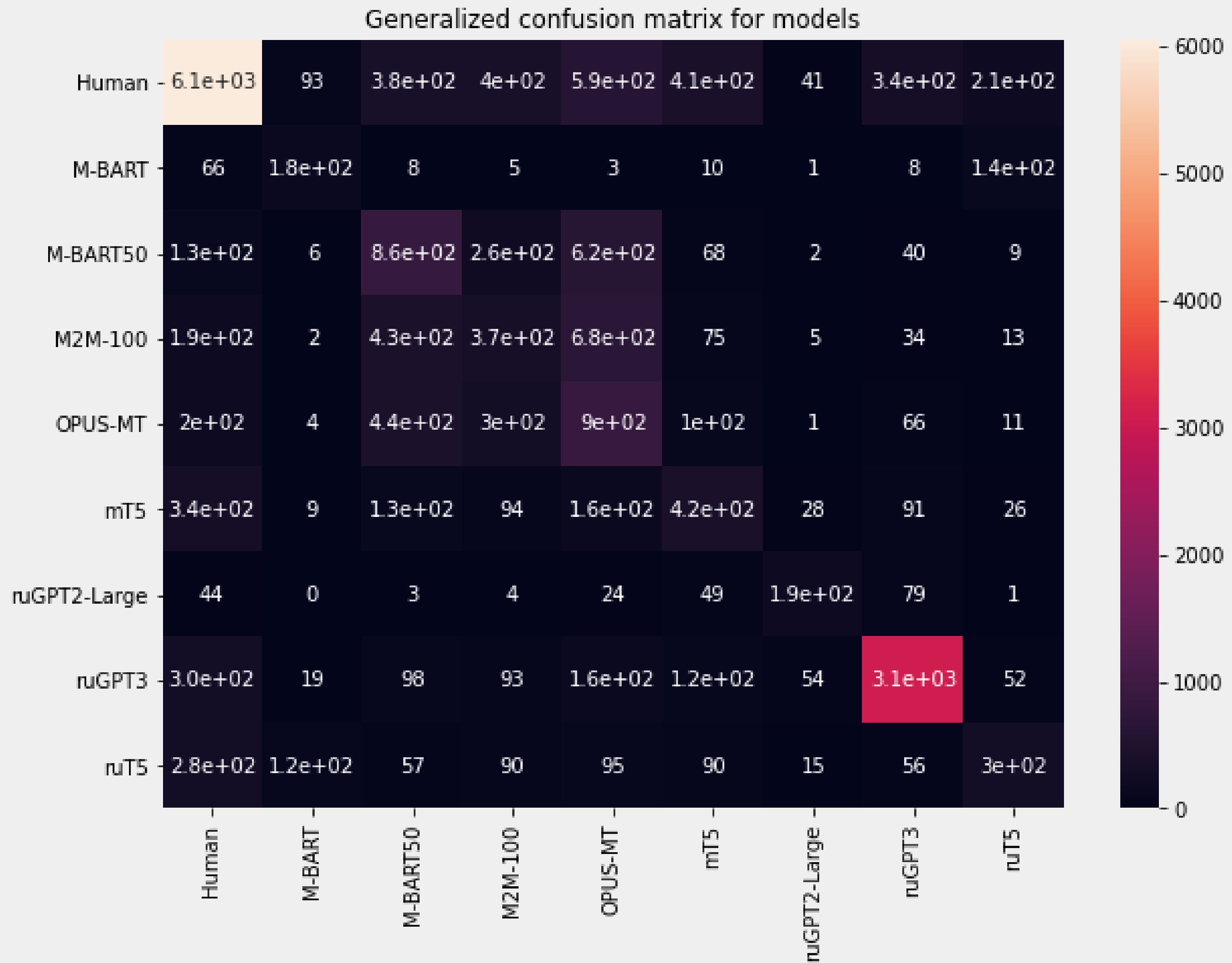
combo model



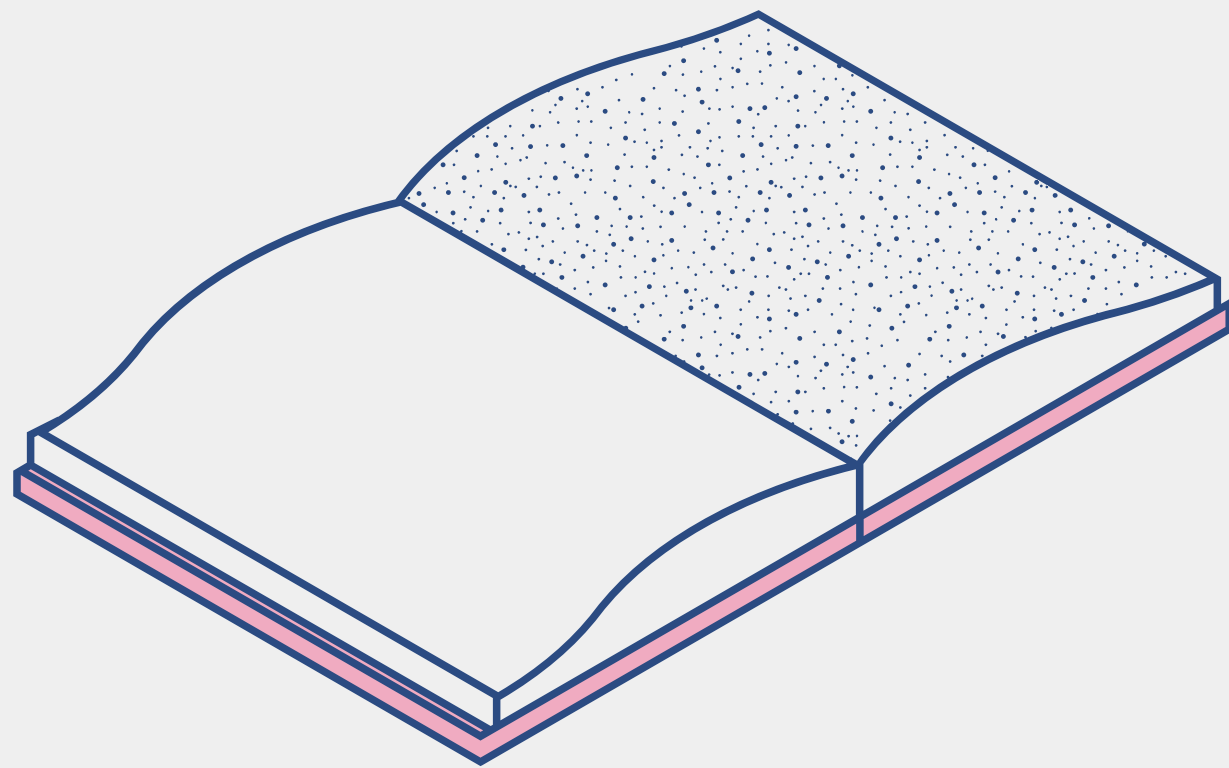
Финальные результаты

RESULTS Combo model with human

	precision	recall	f1-score	support
Human	0.80	0.71	0.75	8524
M-BART	0.42	0.43	0.42	418
M-BART50	0.36	0.43	0.39	1986
M2M-100	0.23	0.21	0.22	1804
OPUS-MT	0.28	0.45	0.34	2014
mT5-Large	0.12	0.11	0.12	810
mT5-Small	0.20	0.24	0.22	490
ruGPT2-Large	0.57	0.48	0.52	395
ruGPT3-Large	0.44	0.52	0.48	1645
ruGPT3-Medium	0.48	0.22	0.30	1170
ruGPT3-Small	0.44	0.50	0.47	1155
ruT5-Base	0.43	0.57	0.49	440
ruT5-Base-Multitask	0.28	0.02	0.04	370
ruT5-Large	0.06	0.03	0.04	290
accuracy			0.50	21511
macro avg	0.36	0.35	0.34	21511
weighted avg	0.52	0.50	0.50	21511

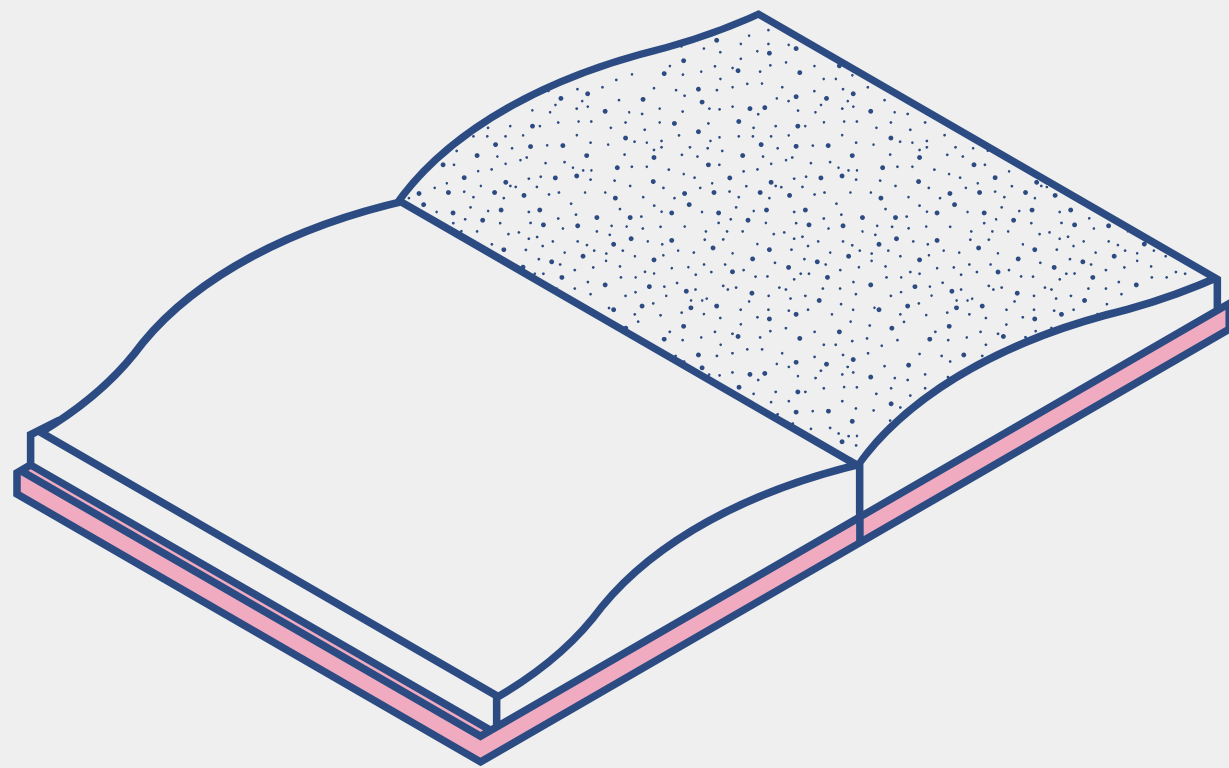


Проблемы



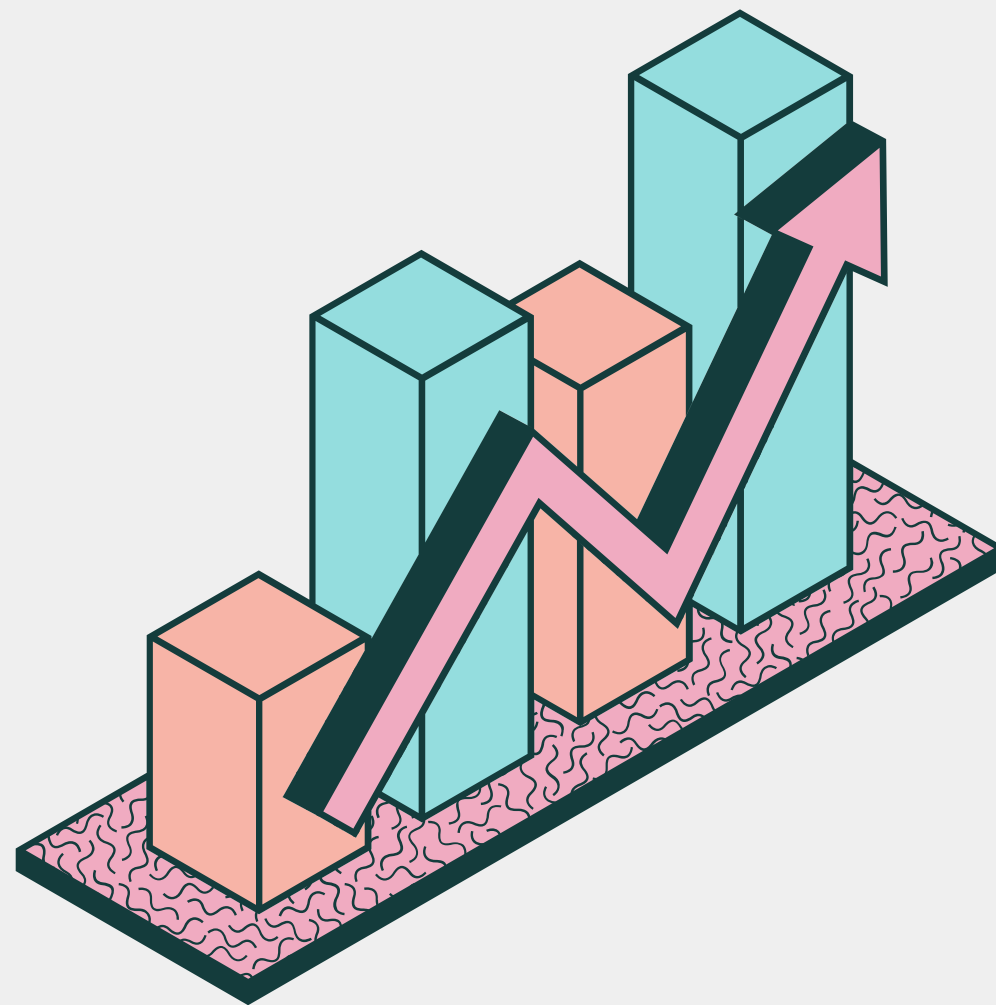
- большой объем данных
- требуется много ресурсов и времени для обработки и обучения моделей
- дисбаланс классов в мультиклассовой задаче

Предложения



- использовать модели для генерации данных, чтобы сбалансировать датасет
- лучше подобрать гиперпараметры для BERT, использовать модель большего размера

Выводы



- Можно хорошо отличать человека от модели
- Можно использовать мультиклассификацию как оценку генерации
- При этом модели разного размера, но одной архитектуры часто путаются, объединение моделей разных моделей дало +0.08 к аккьюраси
- Мультилингвальные модели больше похоже между собой, чем модели одной архитектуры (например, mT5 и ruT5)
- Мультилингвальные модели (M2M100 и mT5) чаще путаются с человеком и хуже определяются (лучше генерируют?)
- Признаков оказывается мало для хорошей классификации, хорошо работают символы и TF-IDF для юниграм, биграмм и триграм. Их совмещение дает небольшой прирост качества

A man with a beard is seen from the side, sitting at a desk and looking at a computer monitor. The entire image has a teal-colored overlay. The man is wearing a light-colored t-shirt. The desk has a computer monitor, a mouse, and some papers. The background is slightly blurred, showing a window and some office equipment.

Литература

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. arXiv preprint arXiv:1910.14659.

Взяли идею использовать маски для различения моделей.

<https://github.com/EkaterinaVoloshina/RuATD>

