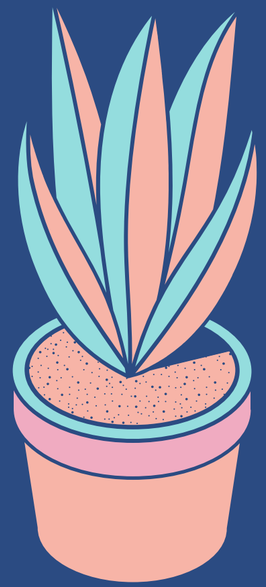
RuATD: распознавание сгенерированных текстов

Катя Волошина & Полина Кудрявцева

Задачи

ПО МОТИВАМ СОРЕВНОВАНИЯ ДИАЛОГ-2022 RUATD

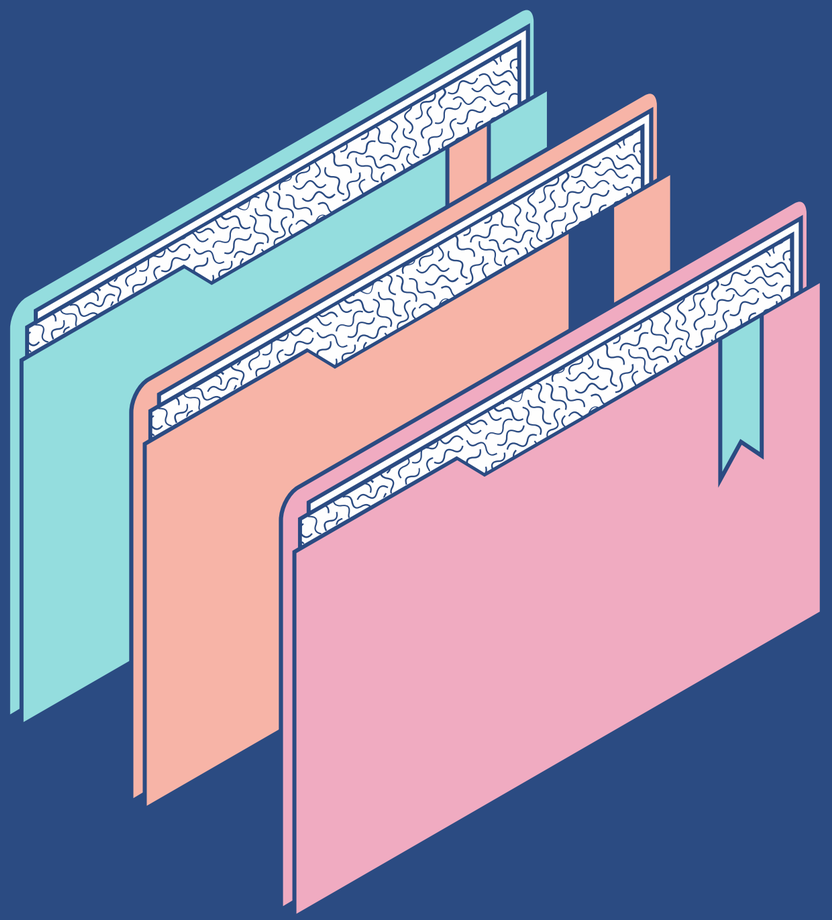


1

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.

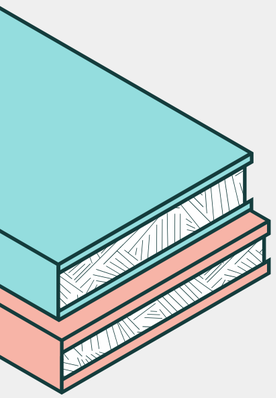
Команда и роли

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

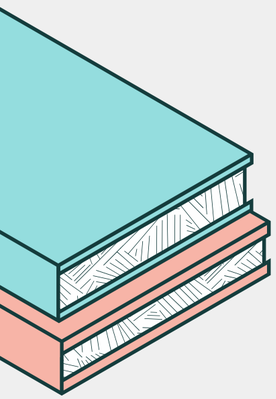


Данные

- тренировочная выборка: 129065 текстов
- тестовая выборка: 64533 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.	H	H
Меня покусали комары.	H	H
Меня похитили муски.	M	Opus-MT
Я был готов помочь ему в опасности своей жизни.	M	Opus-MT
Моя квартира находится меньше чем в пяти минутах пешком от станции.	H	H

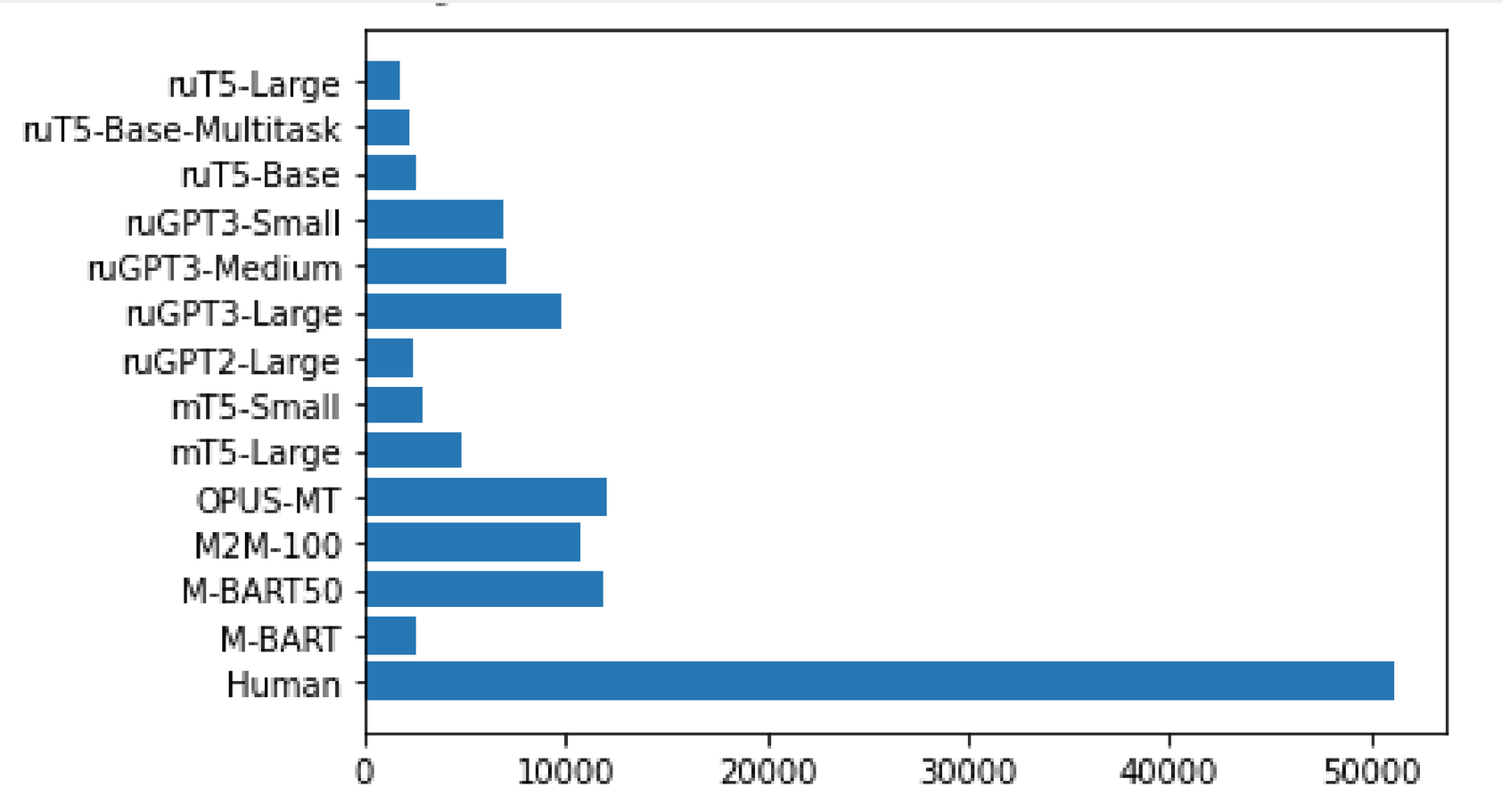


Данные

- 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

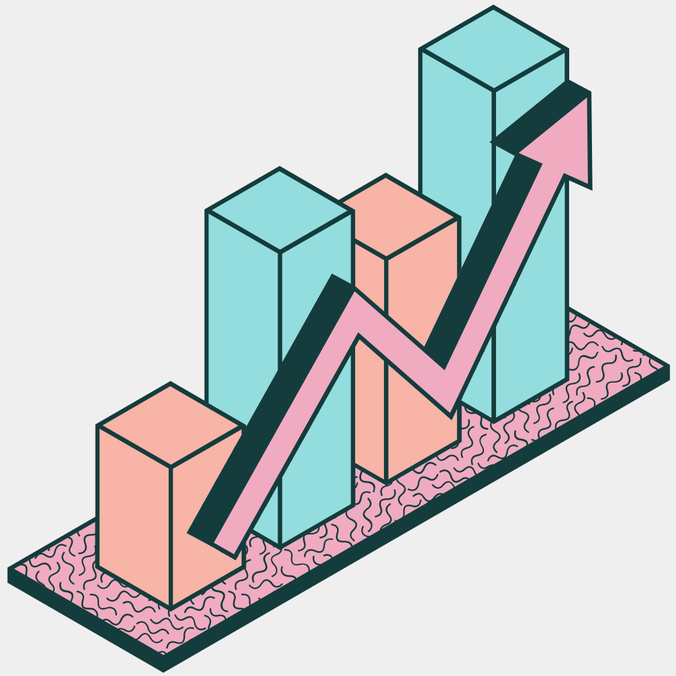
LogReg на TF-IDF

Простейший способ преобразования данных и простейший классификатор

Fine-tuned ruBERT

Это state-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, какова доля правильных ответов у модели

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

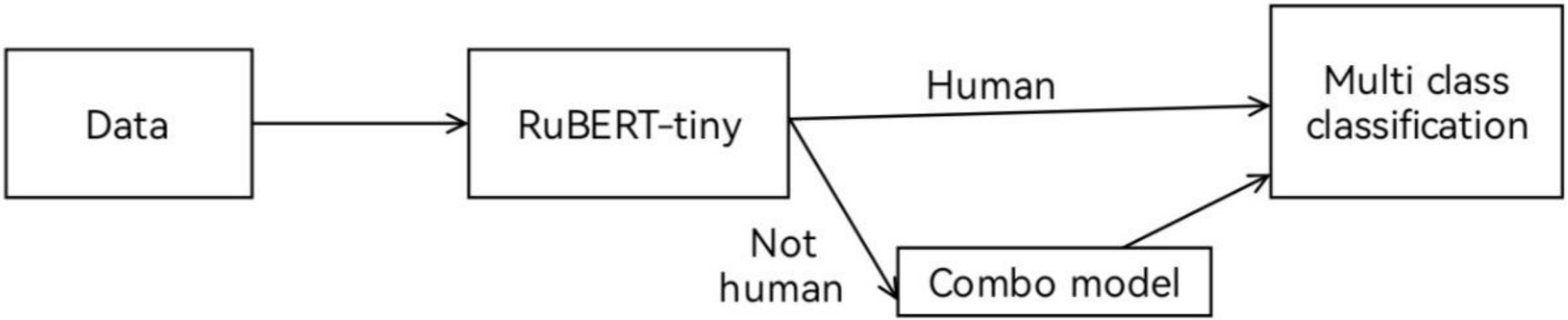
Подготовка данных

- Токенизация, pymorphy2
- Лемматизация, pymorphy2
- Векторизация: TF-IDF, word2vec

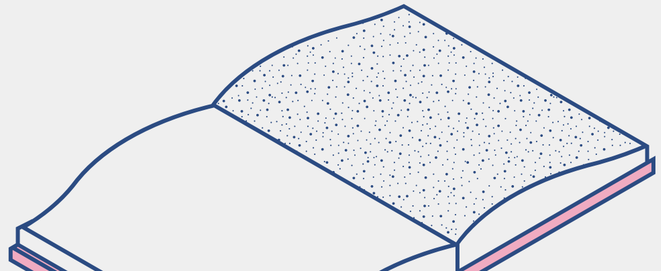
- Выделение признаков у предложений:
  - длина в символах и в токенах
  - кол-во гласных на символы
  - кол-во пробелов на символы
  - кол-во пунктуационных знаков на символы
  - средняя длина слов
  - кол-во длинных слов (>= 10 букв) на все слова
  - кол-во коротких слов (<=3 букв) на все слова
  - readability:
    - Dale-Chall
    - Gunning-Fog
    - Flesch

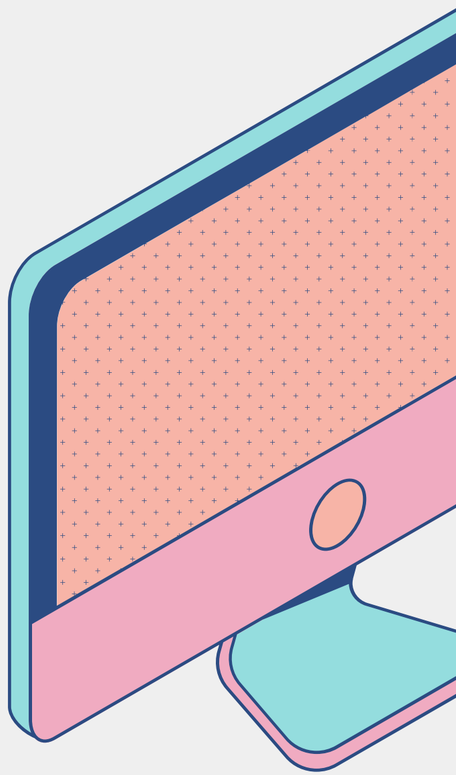
- морфология:
  - кол-во служебных частей речи (ч.р.) на все слова
  - кол-во самостоятельных ч.р. на все слова
  - кол-во именных ч.р. на все слова
  - кол-во глагольных ч.р.на все слова
  - отношение кол-ва именных ч.р. к глагольным

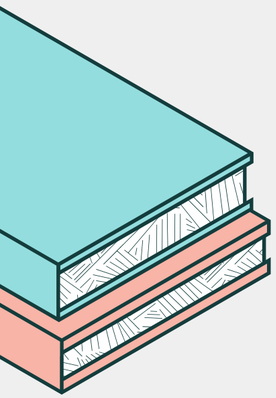
Двухэтапный пайплайн



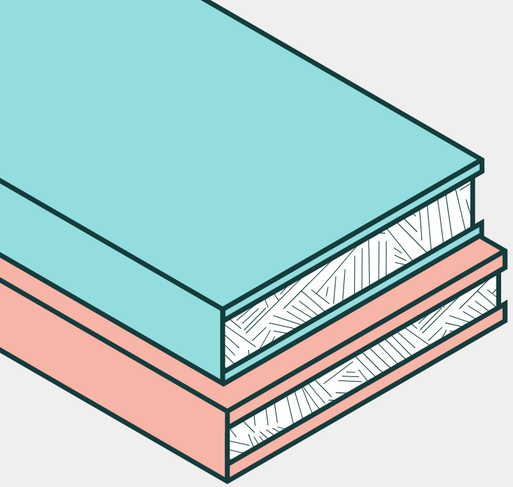
Почему мы выбрали RuBERT-tiny?







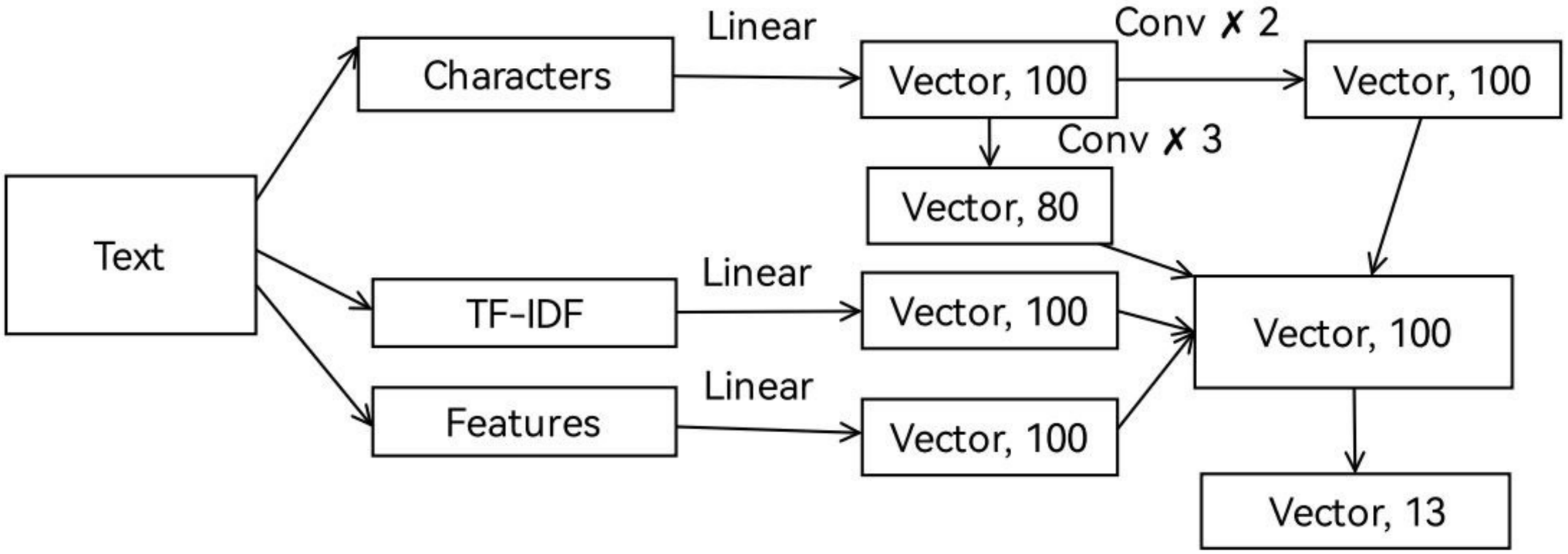
encoding	classifier	binary: accuracy
BoW+TF-IDF	FNN	0.73
word2vec	LSTM	0.59
character	CNN	0,66
Native LM	fine-tuning	**0.81**
text features	FNN	0.6
text features	Logreg	0.61
text features	RandomForest	0.64
BERT		0.79622
TF-DF	Logreg	0.63562



Почему мы выбрали Combo model? И что это вообще?

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
BERT		**0.59914**	
TF-DF	Logreg	**0.44300**	

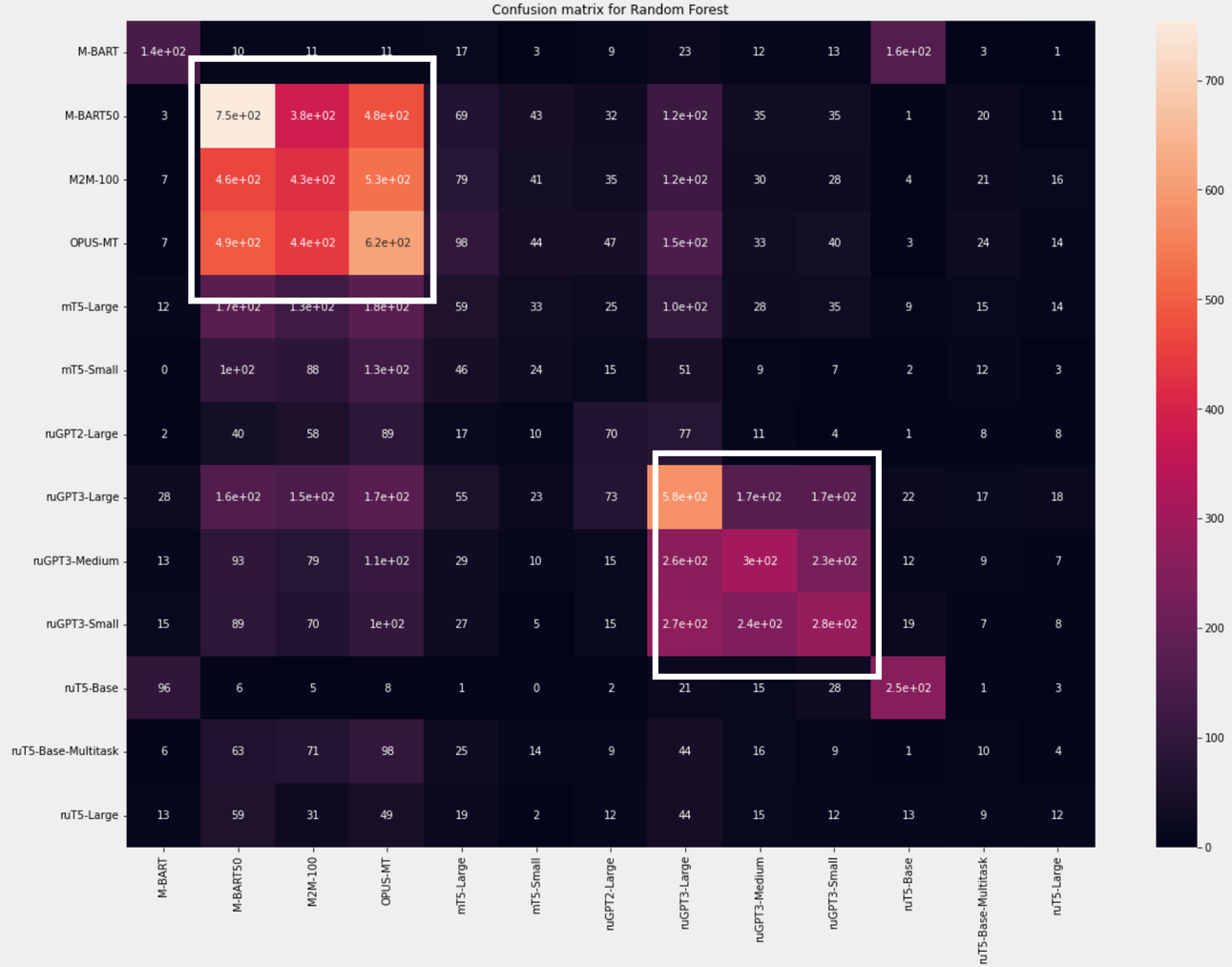
Combo model


Text
Characters
Linear
Vector, 100
Conv ✗ 2
Vector, 100
Conv ✗ 3
Vector, 80
TF-IDF
Linear
Vector, 100
Features
Linear
Vector, 100
Vector, 100
Vector, 13

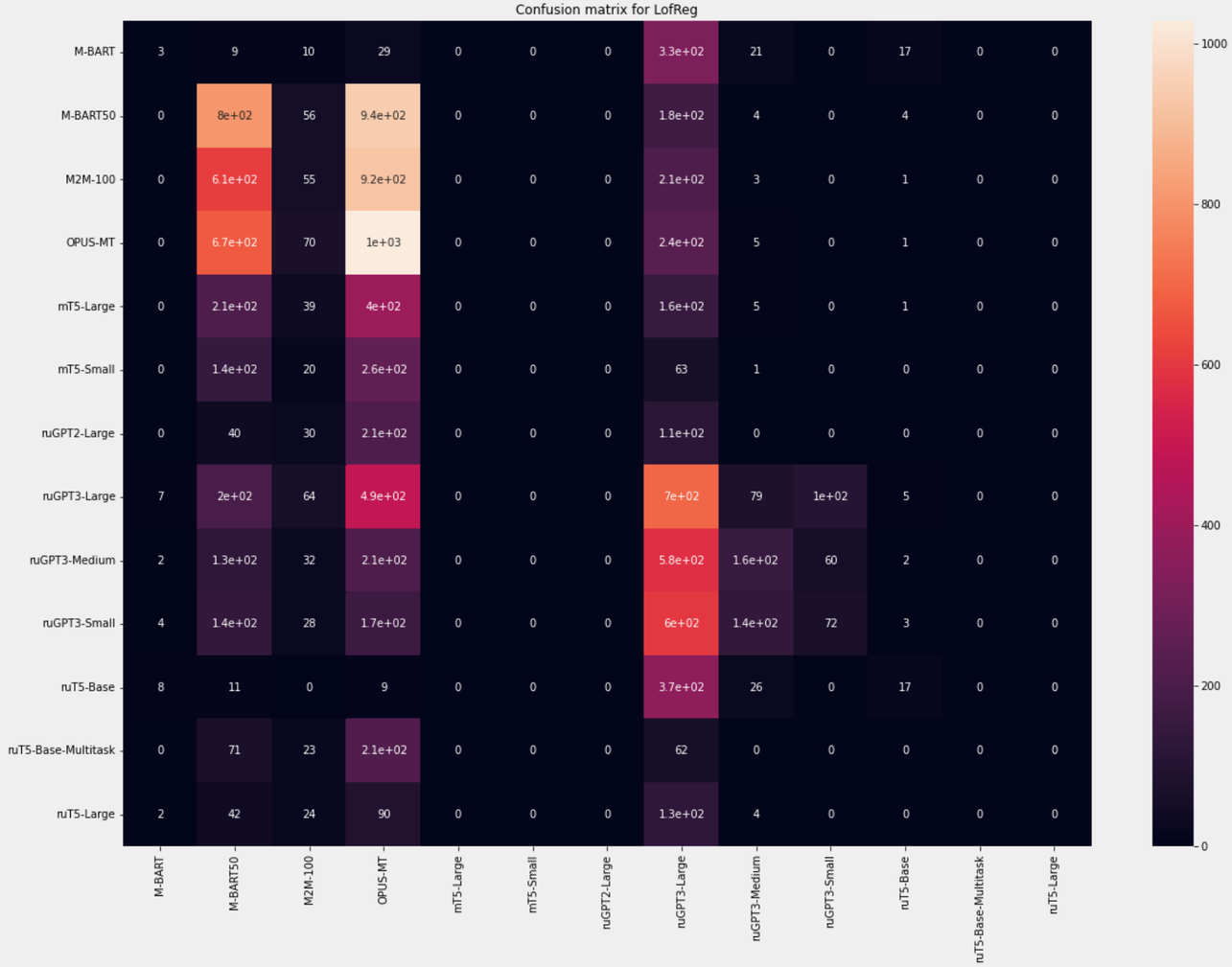
	precision	recall	f1-score	support
M-BART	0.52	0.36	0.42	418
M-BART50	0.41	0.48	0.44	1986
M2M-100	0.29	0.29	0.29	1804
OPUS-MT	0.33	0.41	0.36	2014
mT5-Large	0.25	0.25	0.25	810
mT5-Small	0.29	0.22	0.25	490
ruGPT2-Large	0.64	0.50	0.56	395
ruGPT3-Large	0.48	0.60	0.53	1645
ruGPT3-Medium	0.45	0.10	0.16	1170
ruGPT3-Small	0.46	0.59	0.52	1155
ruT5-Base	0.54	0.72	0.61	440
ruT5-Base-Multitask	0.38	0.02	0.05	370
ruT5-Large	0.15	0.06	0.08	290
accuracy			0.39	12987
macro avg	0.40	0.35	0.35	12987
weighted avg	0.39	0.39	0.37	12987

FNN

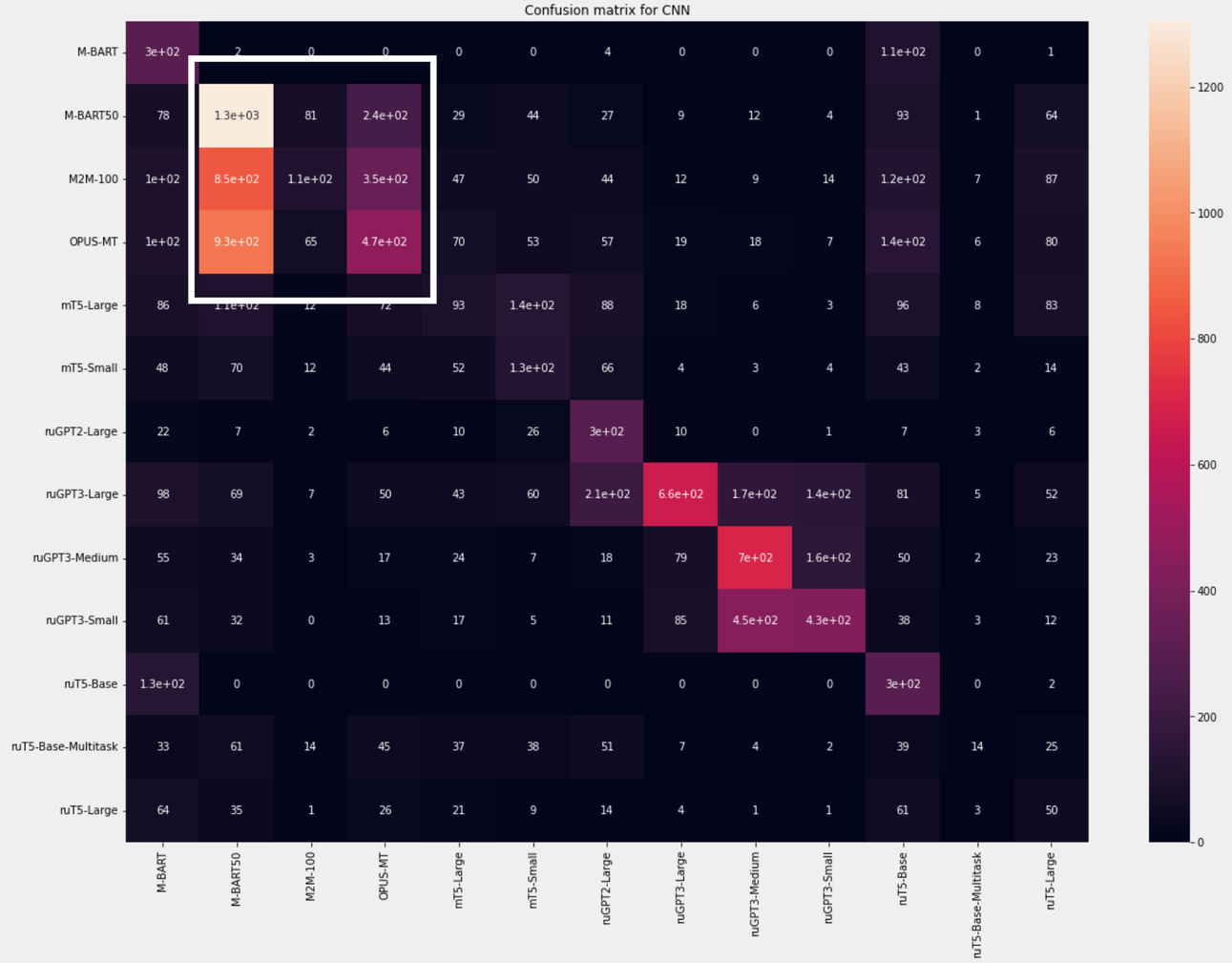
Random Forest



LogReg


Confusion matrix for LofReg
M-BART 3 9 10 29 0 0 0 3.3e+02 21 0 17 0 0
M-BART50 0 8e+02 56 9.4e+02 0 0 0 1.8e+02 4 0 4 0 0
M2M-100 0 6.1e+02 55 9.2e+02 0 0 0 2.1e+02 3 0 1 0 0
OPUS-MT 0 6.7e+02 70 1e+03 0 0 0 2.4e+02 5 0 1 0 0
mT5-Large 0 2.1e+02 39 4e+02 0 0 0 1.6e+02 5 0 1 0 0
mT5-Small 0 1.4e+02 20 2.6e+02 0 0 0 63 1 0 0 0 0
ruGPT2-Large 0 40 30 2.1e+02 0 0 0 1.1e+02 0 0 0 0 0
ruGPT3-Large 7 2e+02 64 4.9e+02 0 0 0 7e+02 79 1e+02 5 0 0
ruGPT3-Medium 2 1.3e+02 32 2.1e+02 0 0 0 5.8e+02 1.6e+02 60 2 0 0
ruGPT3-Small 4 1.4e+02 28 1.7e+02 0 0 0 6e+02 1.4e+02 72 3 0 0
ruT5-Base 8 11 0 9 0 0 0 3.7e+02 26 0 17 0 0
ruT5-Base-Multitask 0 71 23 2.1e+02 0 0 0 62 0 0 0 0 0
ruT5-Large 2 42 24 90 0 0 0 1.3e+02 4 0 0 0 0
M-BART M-BART50 M2M-100 OPUS-MT mT5-Large mT5-Small ruGPT2-Large ruGPT3-Large ruGPT3-Medium ruGPT3-Small ruT5-Base ruT5-Base-Multitask ruT5-Large
1000 800 600 400 200 0

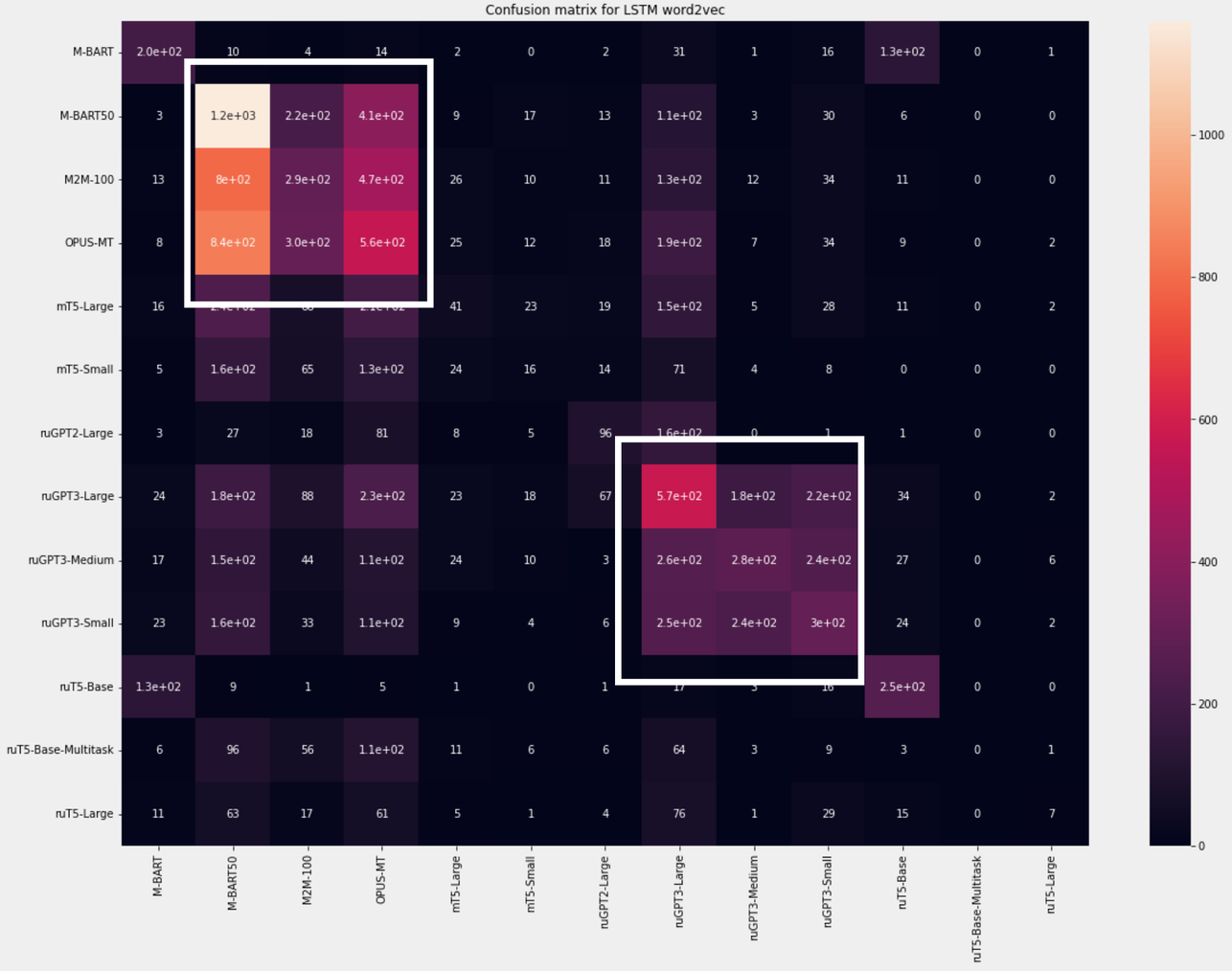
CNN


Confusion matrix for CNN
M-BART 3e+02 2 0 0 0 0 4 0 0 0 1.1e+02 0 1
M-BART50 78 1.3e+03 81 2.4e+02 29 44 27 9 12 4 93 1 64
M2M-100 1e+02 8.5e+02 1.1e+02 3.5e+02 47 50 44 12 9 14 1.2e+02 7 87
OPUS-MT 1e+02 9.3e+02 65 4.7e+02 70 53 57 19 18 7 1.4e+02 6 80
mT5-Large 86 1.1e+02 12 72 93 1.4e+02 88 18 6 3 96 8 83
mT5-Small 48 70 12 44 52 1.3e+02 66 4 3 4 43 2 14
ruGPT2-Large 22 7 2 6 10 26 3e+02 10 0 1 7 3 6
ruGPT3-Large 98 69 7 50 43 60 2.1e+02 6.6e+02 1.7e+02 1.4e+02 81 5 52
ruGPT3-Medium 55 34 3 17 24 7 18 79 7e+02 1.6e+02 50 2 23
ruGPT3-Small 61 32 0 13 17 5 11 85 4.5e+02 4.3e+02 38 3 12
ruT5-Base 1.3e+02 0 0 0 0 0 0 0 0 0 3e+02 0 2
ruT5-Base-Multitask 33 61 14 45 37 38 51 7 4 2 39 14 25
ruT5-Large 64 35 1 26 21 9 14 4 1 1 61 3 50
M-BART M-BART50 M2M-100 OPUS-MT mT5-Large mT5-Small ruGPT2-Large ruGPT3-Large ruGPT3-Medium ruGPT3-Small ruT5-Base ruT5-Base-Multitask ruT5-Large
1200 1000 800 600 400 200 0

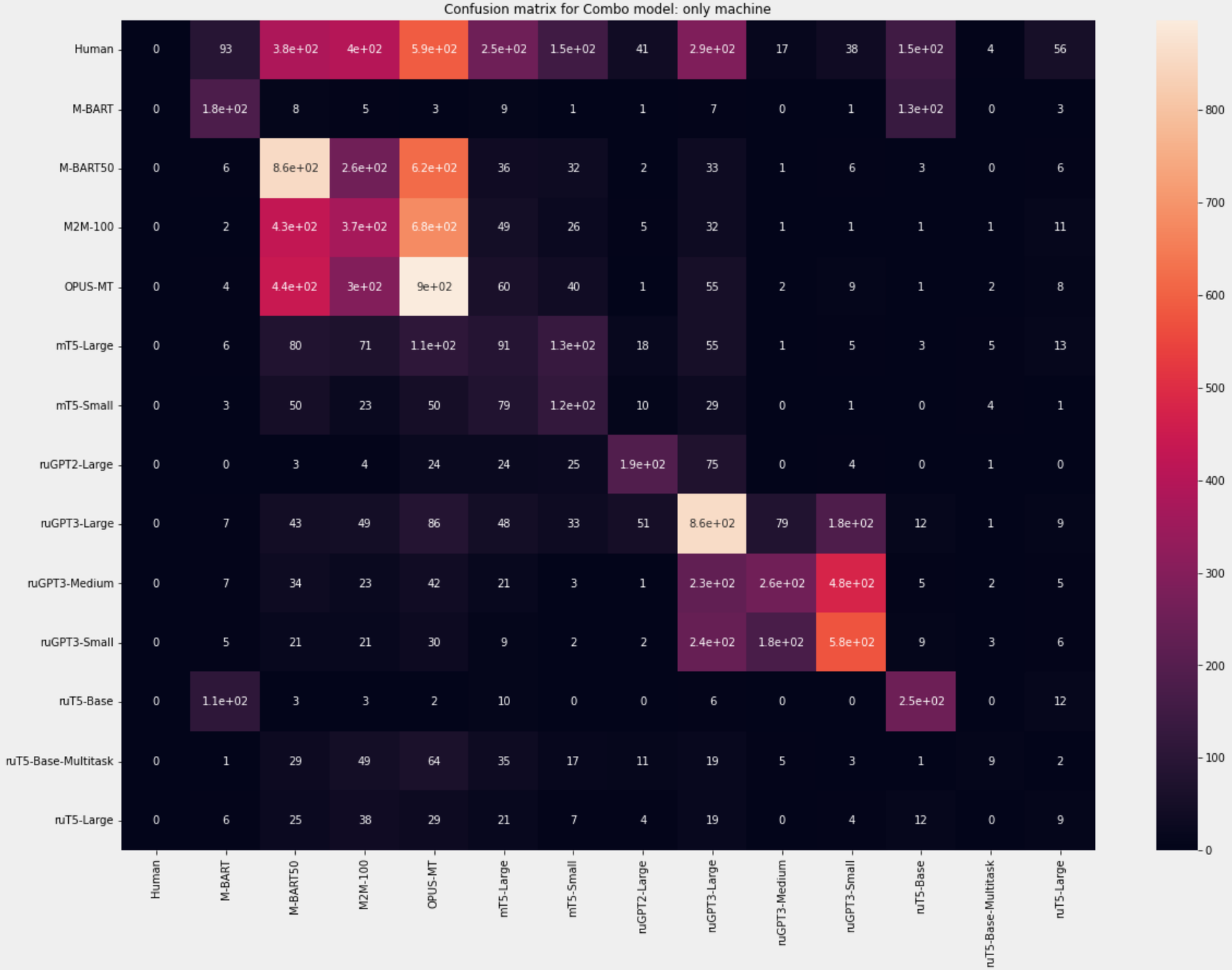
LSTM

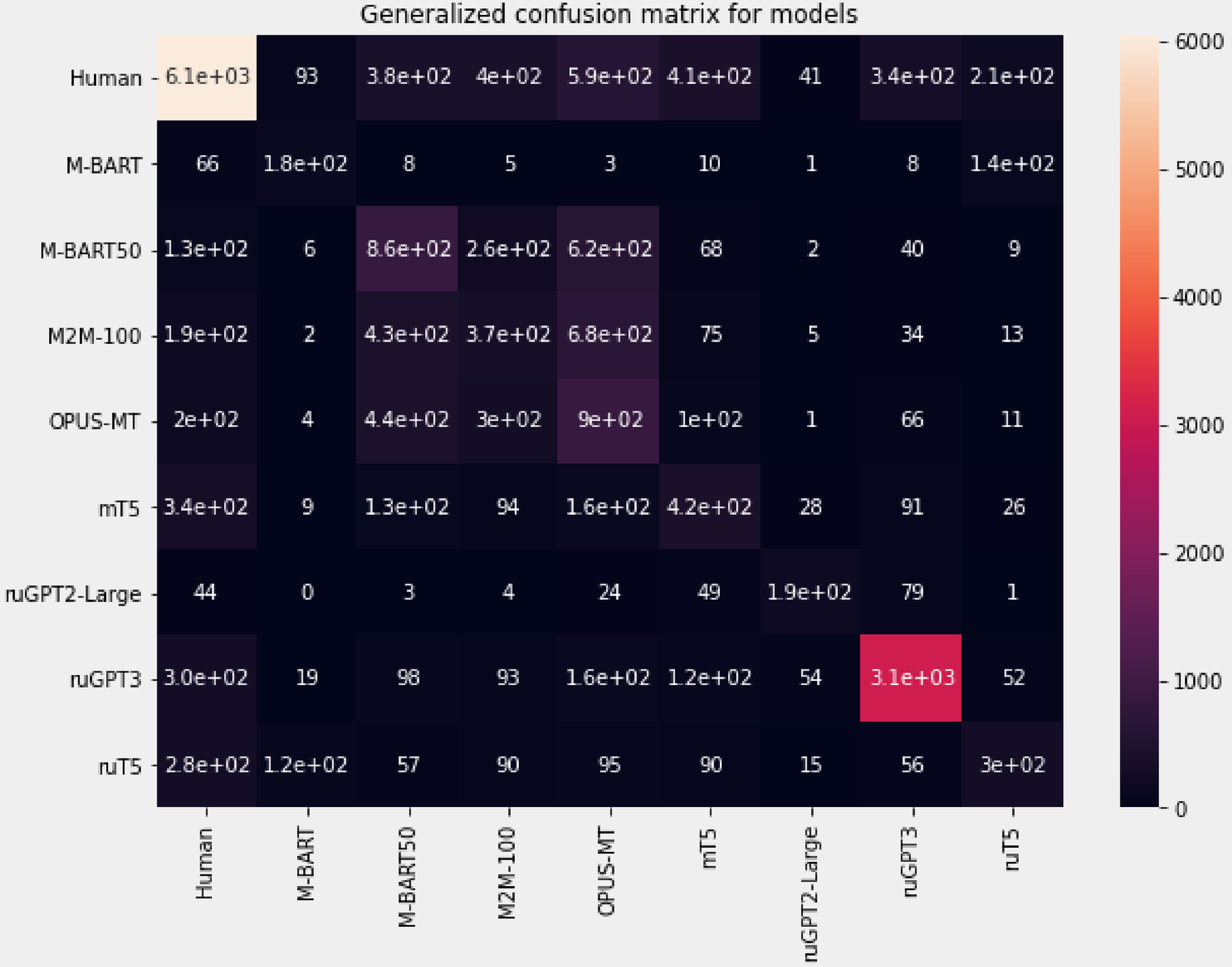
TF-IDF

word2vec

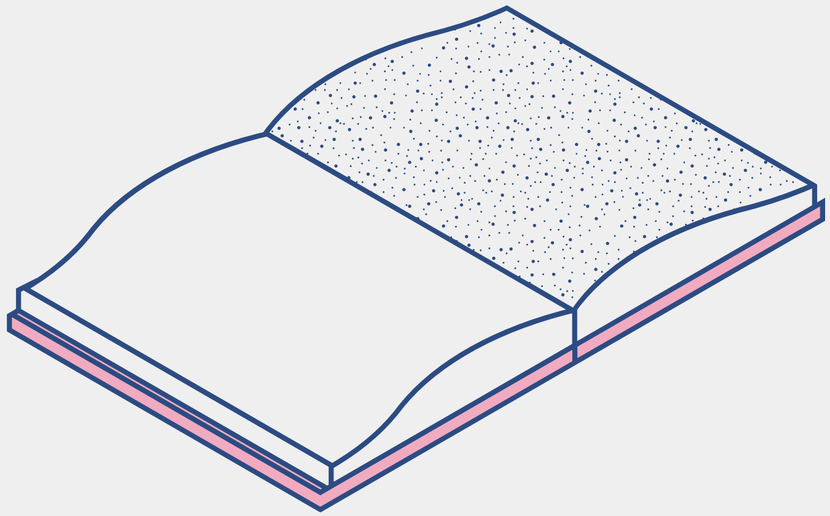


combo model


Confusion matrix for Combo model: only machine
Human 0 93 3.8e+02 4e+02 5.9e+02 2.5e+02 1.5e+02 41 2.9e+02 17 38 1.5e+02 4 56
M-BART 0 1.8e+02 8 5 3 9 1 1 7 0 1 1.3e+02 0 3
M-BART50 0 6 8.6e+02 2.6e+02 6.2e+02 36 32 2 33 1 6 3 0 6
M2M-100 0 2 4.3e+02 3.7e+02 6.8e+02 49 26 5 32 1 1 1 1 11
OPUS-MT 0 4 4.4e+02 3e+02 9e+02 60 40 1 55 2 9 1 2 8
mT5-Large 0 6 80 71 1.1e+02 91 1.3e+02 18 55 1 5 3 5 13
mT5-Small 0 3 50 23 50 79 1.2e+02 10 29 0 1 0 4 1
ruGPT2-Large 0 0 3 4 24 24 25 1.9e+02 75 0 4 0 1 0
ruGPT3-Large 0 7 43 49 86 48 33 51 8.6e+02 79 1.8e+02 12 1 9
ruGPT3-Medium 0 7 34 23 42 21 3 1 2.3e+02 2.6e+02 4.8e+02 5 2 5
ruGPT3-Small 0 5 21 21 30 9 2 2 2.4e+02 1.8e+02 5.8e+02 9 3 6
ruT5-Base 0 1.1e+02 3 3 2 10 0 0 6 0 0 2.5e+02 0 12
ruT5-Base-Multitask 0 1 29 49 64 35 17 11 19 5 3 1 9 2
ruT5-Large 0 6 25 38 29 21 7 4 19 0 4 12 0 9
Human
M-BART
M-BART50
M2M-100
OPUS-MT
mT5-Large
mT5-Small
ruGPT2-Large
ruGPT3-Large
ruGPT3-Medium
ruGPT3-Small
ruT5-Base
ruT5-Base-Multitask
ruT5-Large
800
700
600
500
400
300
200
100
0

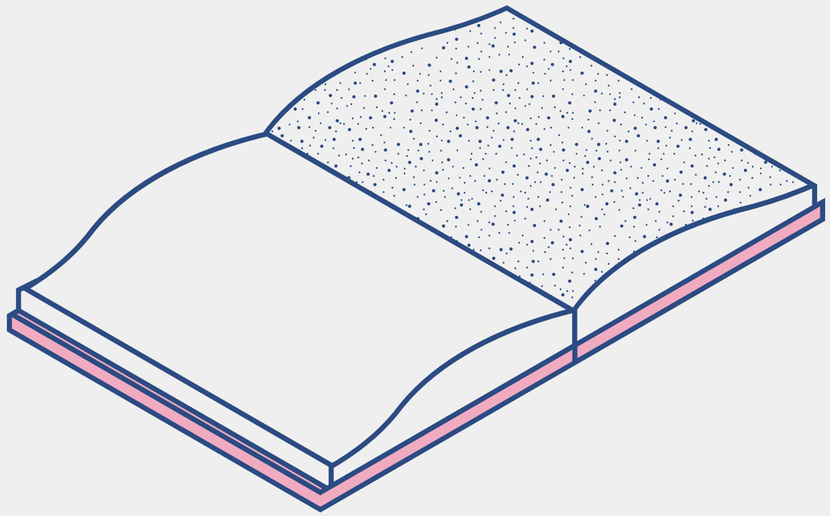

Generalized confusion matrix for models
Human
M-BART
M-BART50
M2M-100
OPUS-MT
mT5
ruGPT2-Large
ruGPT3
ruT5
6.1e+03 93 3.8e+02 4e+02 5.9e+02 4.1e+02 41 3.4e+02 2.1e+02
66 1.8e+02 8 5 3 10 1 8 1.4e+02
1.3e+02 6 8.6e+02 2.6e+02 6.2e+02 68 2 40 9
1.9e+02 2 4.3e+02 3.7e+02 6.8e+02 75 5 34 13
2e+02 4 4.4e+02 3e+02 9e+02 1e+02 1 66 11
3.4e+02 9 1.3e+02 94 1.6e+02 4.2e+02 28 91 26
44 0 3 4 24 49 1.9e+02 79 1
3.0e+02 19 98 93 1.6e+02 1.2e+02 54 3.1e+03 52
2.8e+02 1.2e+02 57 90 95 90 15 56 3e+02
6000
5000
4000
3000
2000
1000
0

Проблемы



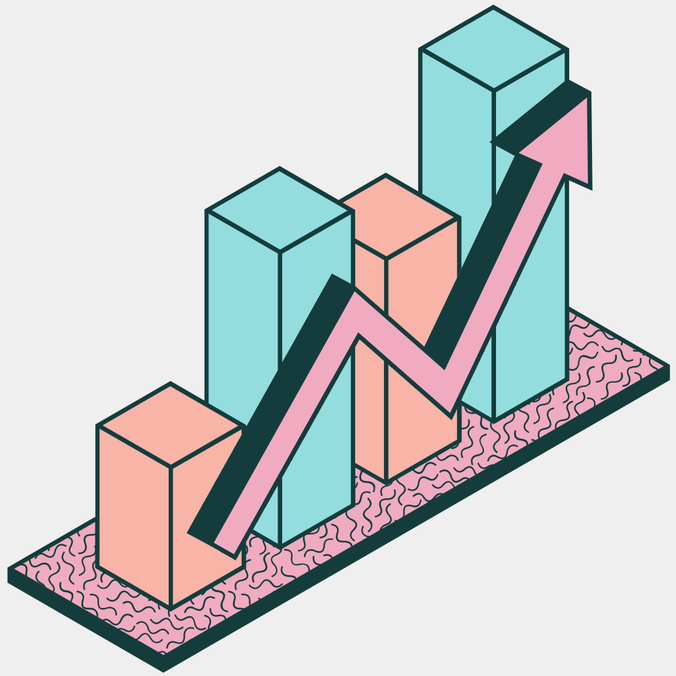
- большой объем данных
- требуется много ресурсов и времени для обработки и обучения моделей
- дисбаланс классов в мультиклассовой задаче

Предложения

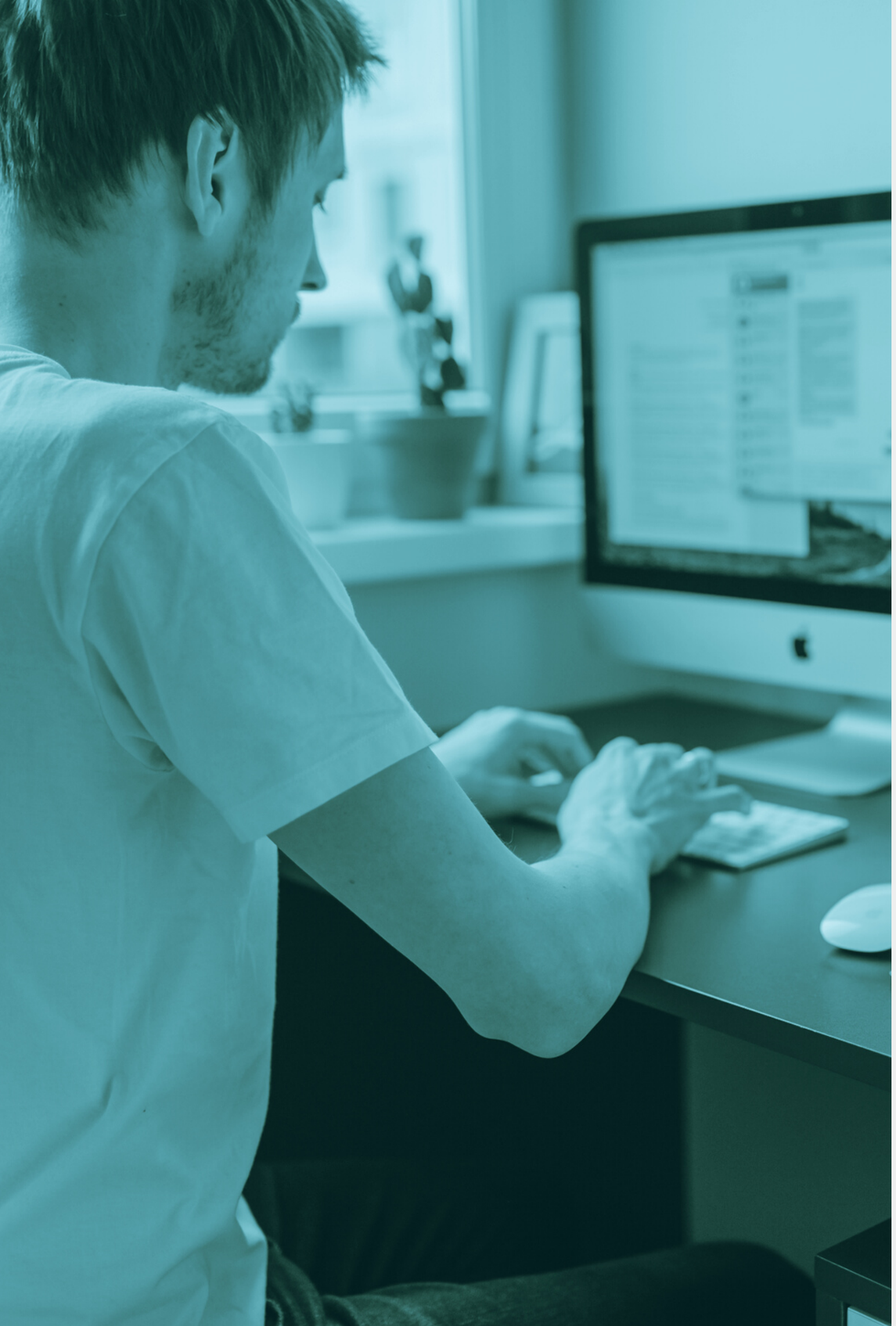


- использовать модели для генерации данных, чтобы сбалансировать датасет
- лучше подобрать гиперпараметры для BERT, использовать модель большего размера

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, биграм и триграм. Их совмещение дает небольшой прирост качества



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