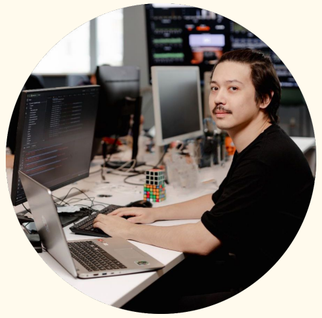
SKILLFACTORY

Хакатон MIFIML

Итоговая презентация

Команда
Gradient Seekers



Владимир Бек

Анастасия Игнатьева

Артём Мартынов

Илья Серов

Евгения Смолякова

Анастасия Яровикова

Задача

Бриф:

Название: Тематическая классификация текстов

Партнёр: Компания «Норси-Транс»

Стэк проекта:

Python, PyTorch, Docker, HTML, CSS, JavaScript, Bootstrap, Flask

Цель:

Создать систему, которая определяет вероятность принадлежности текста к одной или нескольким тематикам из заданного списка

Задачи:

1. Разработать и обучить модель машинного обучения, способную анализировать текст и предсказывать вероятности его принадлежности к одной или нескольким темам
2. Создать пользовательский интерфейс, который принимает текст на вход, отправляет его на классификацию и возвращает категорию, к которой относится текст

Описание решения

Основные этапы

1. EDA, предобработка данных (очистка, токенизация)
2. Оценка распределения данных по классам, проверка кластеризации
3. Классификация методами классического МО
4. single-label классификация с помощью RuBERT модели
5. multi-label классификация с помощью RuBERT модели
6. Предсказание на тестовых данных
7. Разработка UI интерфейса и Flask приложения
8. Тестирование

Описание решения

1. EDA, предобработка данных (очистка, токенизация)

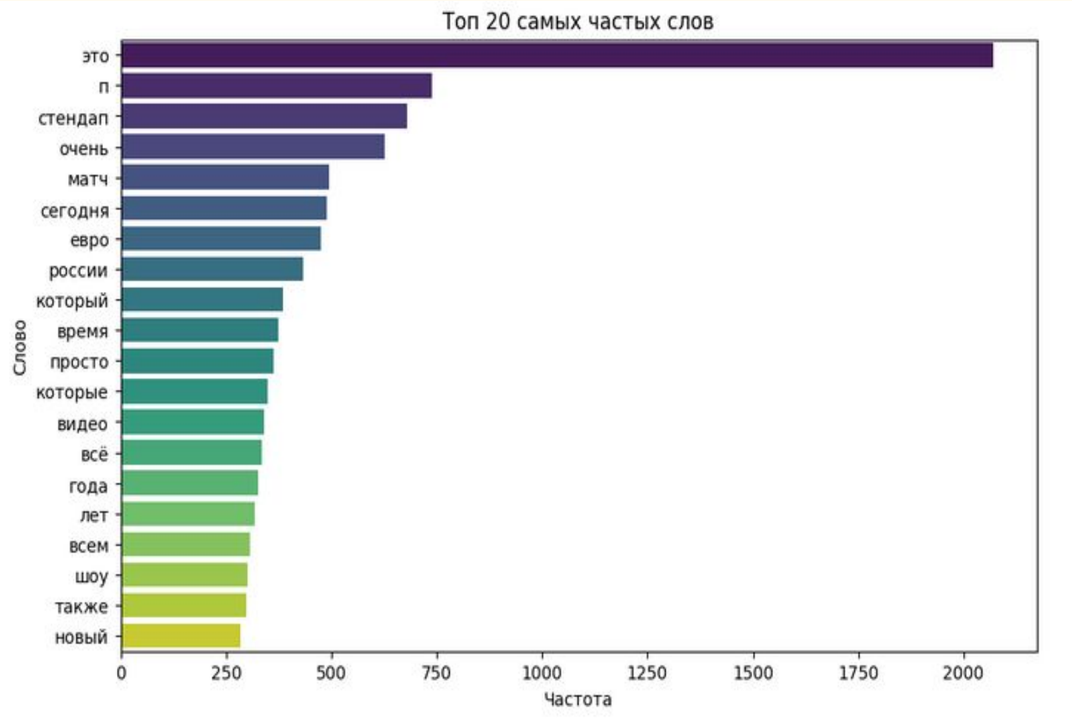
- Анализ данных
- Очистка текста
- Токенизация (подготовка к моделированию):
  - обработка/токенизация + векторизация + модель

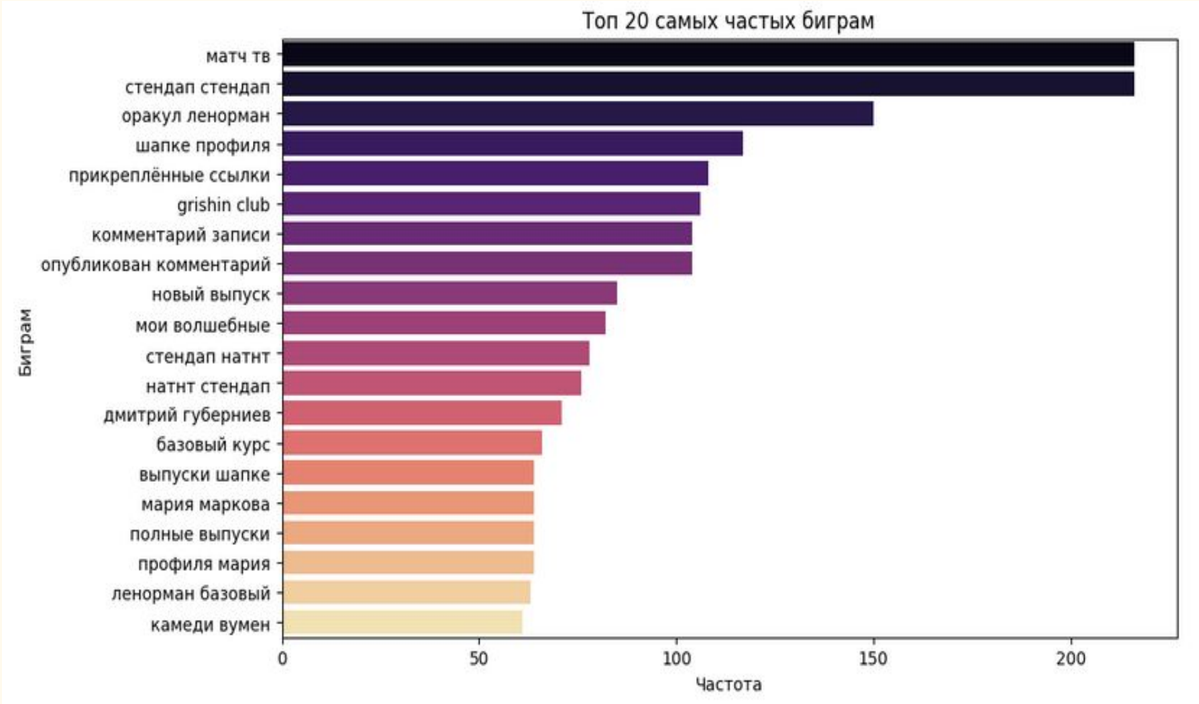
Наименование переменной/-ых	Статус
['doc_text', 'image2text', 'speech2text']	Исходные данные, из переменной doc_text убраны emoji
[‘emojis_doc_text’]	Текстовые описания (на английском) оригинальных эмоций из doc_text
['cleaned_doc_text', ‘cleaned_image2text', ‘cleaned_speech2text']	Очищенные исходные переменные
['stemmed_doc_text', 'stemmed_image2text', 'stemmed_speech2text']	Стемматизированные очищенные переменные
['SpaCy_doc_text', 'SpaCy_image2text', 'SpaCy_speech2text']	Очищенные переменные для дальнейшей токенизации с помощью SpaCy

Описание решения

1. Предобработка данных (очистка, токенизация), EDA

1.1 Проверка уникальности слов/биграм

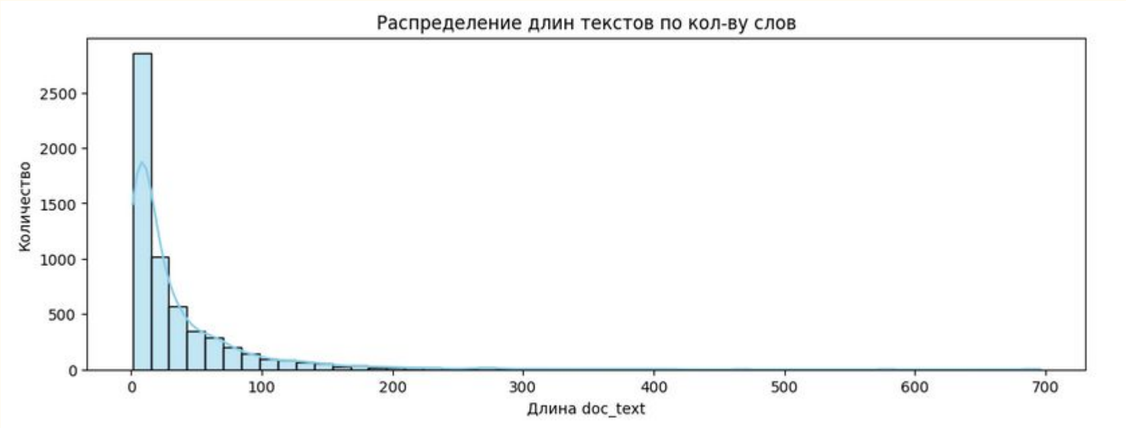


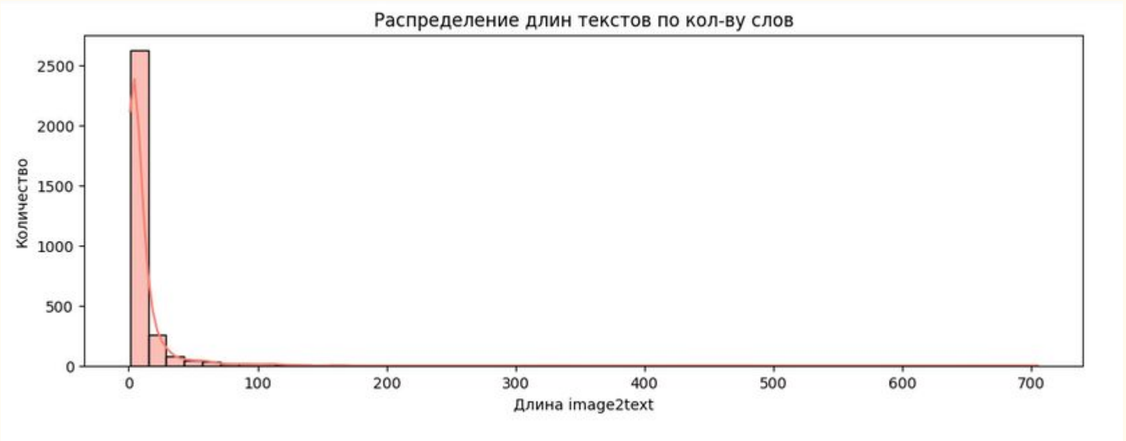


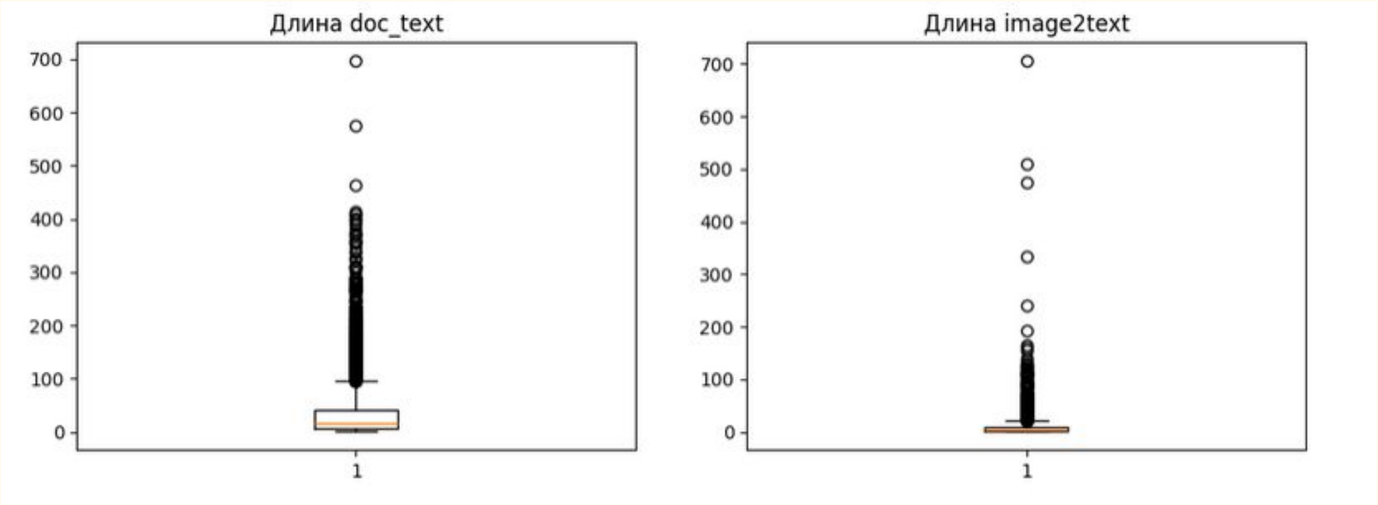
Описание решения

1. Предобработка данных (очистка, токенизация), EDA

1.2 Проверка распределения длин текстов



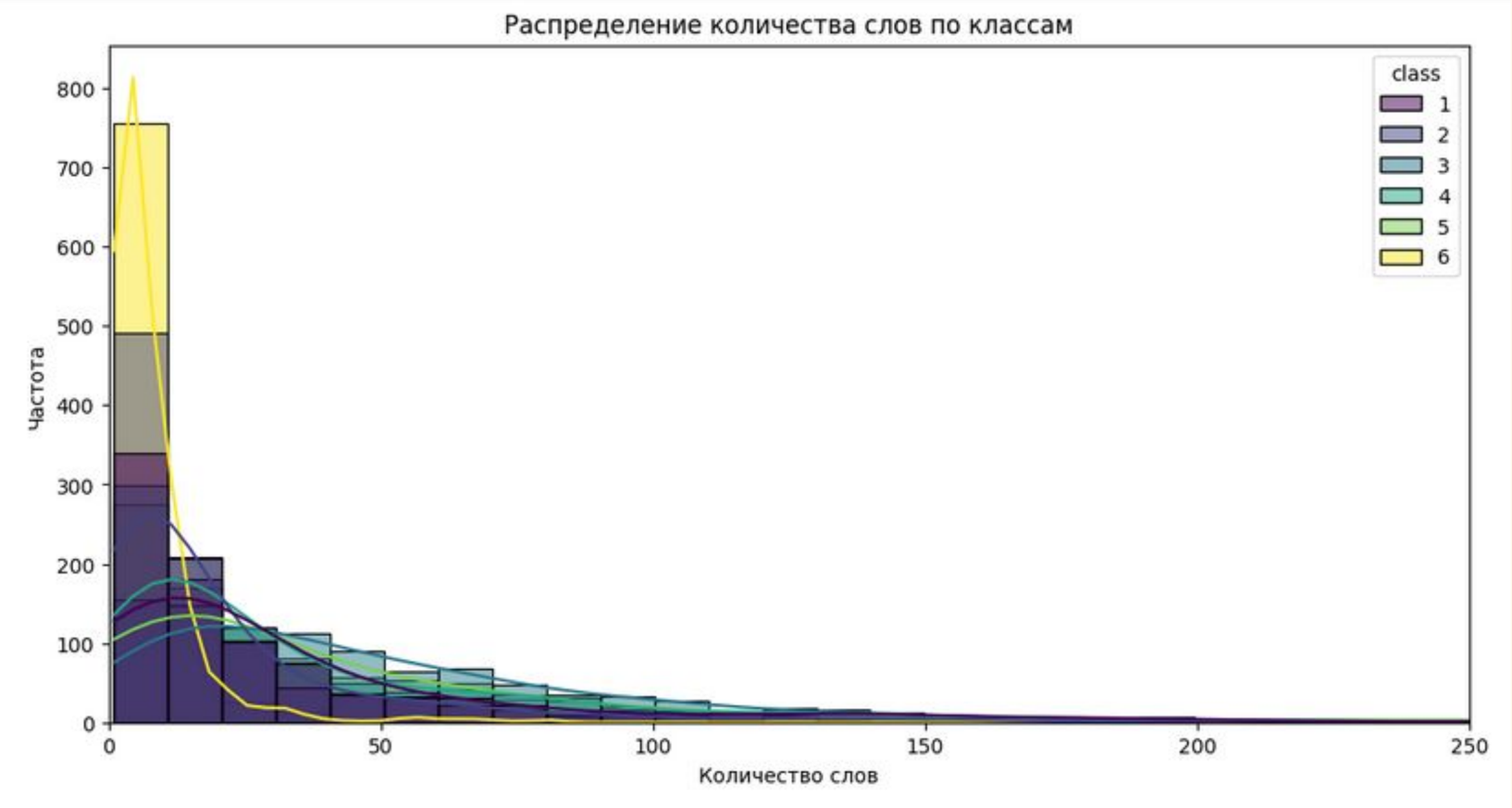




Описание решения

1. Предобработка данных (очистка, токенизация), EDA

1.2 Проверка распределения длин текстов р разрезе классов



Класс	Наименование
1	соцсети
2	личная жизнь
3	политика
4	реклама
5	спорт
6	юмор

Описание решения

1. Предобработка данных (очистка, токенизация), EDA

1.2 Проверка распределения длин текстов в разрезе классов (предобработанный для SpaCy doc_text)

```
Класс 1 топ-10 слов:
['выпуск', 'видео', 'это', 'канал', 'новый', 'смотреть', 'моем', 'канале', 'ссылка', 'ссылки']

Класс 2 топ-10 слов:
['опубликован', 'опубликован комментарий', 'комментарий записи', 'это', 'комментарий', 'записи', 'спасибо', 'очень', 'сегодня', 'день']

Класс 3 топ-10 слов:
['это', 'россии', 'которые', 'крокус', 'очень', 'года', 'люди', 'который', 'время', 'сегодня']

Класс 4 топ-10 слов:
['писать', 'это', 'шоу', 'ленорман', 'июля', 'билеты', 'сегодня', 'выпуск', 'оракул', 'новый']

Класс 5 топ-10 слов:
['это', 'евро', 'матч', 'футбол', 'испания', 'англия', 'динамо', 'сегодня', 'пенальти', 'очень']

Класс 6 топ-10 слов:
['стендап', 'стендап стендап', 'камеди', 'вумен', 'камеди вумен', 'тнт', 'натнт', 'стендап натнт', 'натнт стендап', 'шапке']
```

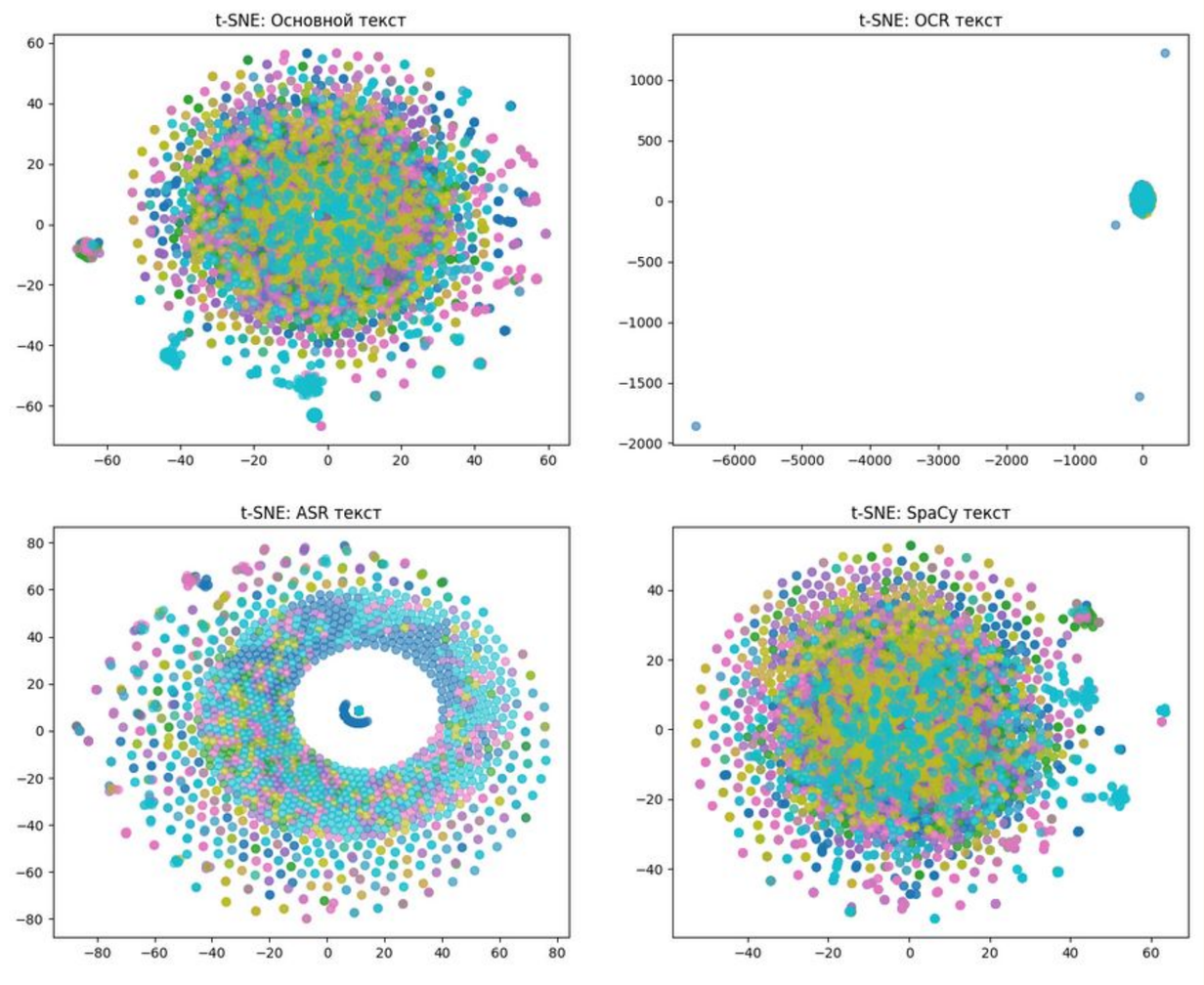
Описание решения

2. Оценка распределения данных по классам, проверка кластеризации

- Предобработка стеммизированных текстов векторизатором TF-IDF
- Визуализация кластеров с помощью t-SNE
- Кластеризация методом k-средних

	Источник	Silhouette	Adjusted Rand
3	SpaCy текст	0.020774	0.015412
0	Основной текст	0.007333	0.002308
1	OCR текст	0.643193	0.000056
2	ASR текст	0.553694	0.000032

- SpaCy/Основной текст: низкие значения (0.01-0.02) → кластеры почти не соответствуют истинным классам
- OCR/ASR: высокий Silhouette (0.5-0.6), но Rand ≈0 → кластеры компактные, но не совпадают с темами

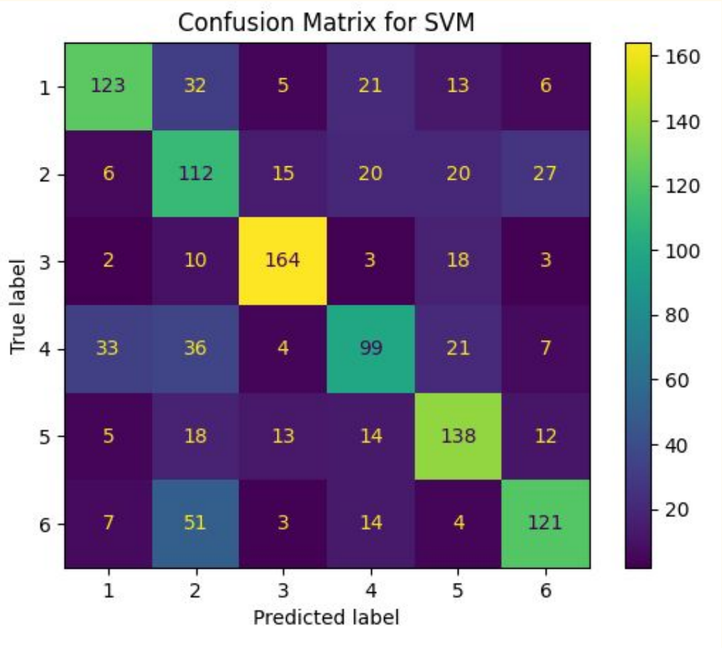


Описание решения

3. Классификация методами классического МО

- Формирование единого датафрейма с развесовкой исходных векторов
- Разделение на тренировочную и тестовую выборки
- Построение моделей

Модель	Precision (macro avg)	F1-score (macro avg)	Accuracy
Logistic Regression	0.64	0.64	0.65
Random Forest	0.64	0.63	0.62
SVM	0.64	0.63	0.63



- Добавление мета-признаков (масштабированы StandardScaler)

Модель	Precision (macro avg)	F1-score (macro avg)	Accuracy
Logistic Regression	0.62	0.61	0.61
Random Forest	0.60	0.59	0.59
SVM	0.53	0.51	0.62

Описание решения

4. single-label классификация с помощью RuBERT модели

- Наименование 'DeepPavlov/rubert-base-cased'
  - лучшее соотношение сложность/качество
- Не требует дополнительной обработки текста перед токенизацией
  - за основу взяты переменные ['doc_text', 'image2text', 'speech2text'] (в переменной ‘doc_text’ уже убраны emoji)
- Добавлена функция, возвращающая дополнительно второй по качеству класс с учётом заданного порога уверенности (метрикой качества оценки выбрана accuracy)
- Построение моделей

Описание решения

4. single-label классификация с помощью RuBERT модели

Epoch	Training Loss	Validation Loss	Accuracy
1	1.007300	0.965336	0.667500
2	0.867800	0.911164	0.665833
3	0.513600	0.980760	0.673333

```
Текст: VladRadimov https://t.me/VladRadimov/950    Наверн...
Предсказания: 5 и 3 (низкая уверенность)
--------------------------------------------------
Текст: 5 сентября в прокат выходит картина «Лгунья» - поб...
Предсказание: 4 (уверенность >= 75%)
--------------------------------------------------
Текст: С праздником, дорогие наши! С Рождеством♡  Череда .
Предсказания: 2 и 4 (низкая уверенность)
--------------------------------------------------
Текст: https://youtu.be/LI0qJ8HR1DI?si=WVcc4q1A96UPEm8S  ...
Предсказание: 1 (уверенность >= 75%)
--------------------------------------------------
Текст: Локомотив!!! Надо еще забивать и выигрывать!!! 💪 n.
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: «Comedy Club» в пятницу в 21:00 на ТНТ 1 9 tht та ...
Предсказание: 6 (уверенность >= 75%)
--------------------------------------------------
Текст: Англия в финале только потому, что её вчера поддер...
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: kruginapole https://t.me/kruginapole/89218 nan жар...
Предсказания: 6 и 2 (низкая уверенность)
--------------------------------------------------
Текст: zarubinreporter https://t.me/zarubinreporter/1990 ...
Предсказание: 3 (уверенность >= 75%)
--------------------------------------------------
Текст: мой утренний шок в Лондоне - поход на стадионный т...
Предсказания: 5 и 2 (низкая уверенность)
--------------------------------------------------
```

Epoch	Training Loss	Validation Loss	Accuracy
1	1.058100	1.002572	0.646667
2	0.928600	0.920146	0.656667
3	0.654300	0.956555	0.663333
4	0.483500	0.959962	0.666667

```
Текст: VladRadimov https://t.me/VladRadimov/950    Наверн...
Предсказания: 3 и 5 (низкая уверенность)
--------------------------------------------------
Текст: 5 сентября в прокат выходит картина «Лгунья» - поб...
Предсказание: 4 (уверенность >= 75%)
--------------------------------------------------
Текст: С праздником, дорогие наши! С Рождеством♡  Череда ...
Предсказания: 2 и 4 (низкая уверенность)
--------------------------------------------------
Текст: https://youtu.be/LI0qJ8HR1DI?si=WVcc4q1A96UPEm8S  ...
Предсказание: 1 (уверенность >= 75%)
--------------------------------------------------
Текст: Локомотив!!! Надо еще забивать и выигрывать!!! 💪 n...
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: «Comedy Club» в пятницу в 21:00 на ТНТ 1 9 tht та ...
Предсказание: 6 (уверенность >= 75%)
--------------------------------------------------
Текст: Англия в финале только потому, что её вчера поддер...
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: kruginapole https://t.me/kruginapole/89218 nan жар...
Предсказания: 2 и 6 (низкая уверенность)
--------------------------------------------------
Текст: zarubinreporter https://t.me/zarubinreporter/1990 ...
Предсказание: 3 (уверенность >= 75%)
--------------------------------------------------
Текст: мой утренний шок в Лондоне - поход на стадионный т...
Предсказания: 2 и 5 (низкая уверенность)
--------------------------------------------------
```

Epoch	Training Loss	Validation Loss	Accuracy
1	0.287200	1.071947	0.656667
2	0.214500	1.192976	0.653333

```
Текст:  Мы снова играем в Кубке России!  Будет большой пр...
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: Что уже сделал Рэтклифф в МЮ, четыре новичка на пр...
Предсказания: 5 и 3 (низкая уверенность)
--------------------------------------------------
Текст: Печальные новости тоже есть…. Утром я проснулась б...
Предсказание: 2 (уверенность >= 75%)
--------------------------------------------------
Текст: ОООООЙ! ЯРСАБАЛЬ! 2-1,   ...
Предсказание: 5 (уверенность >= 75%)
--------------------------------------------------
Текст: У людей разные лингвистические способности, разная...
Предсказание: 2 (уверенность >= 75%)
--------------------------------------------------
```

Описание решения

5. Multi-Bert

- **Данные**

Объединение текстовых полей в единое поле text

Распределение классов: сбалансированное — по 1000 на класс

- **Предобработка**

Удаление стоп-слов, эмоджи, URL, цифр

Стемминг (Snowball Stemmer для русского языка)

Группировка текстов и бинаризация меток (MultiLabelBinarizer)

- **Модель**

Архитектура: BertForSequenceClassification

Предобученная модель: DeepPavlov/rubert-base-cased

Параметры:

max_len = 512

batch_size = 8 (CPU) / 16 (GPU)

epochs = 3

Loss: Binary Cross-Entropy

- **Метрики**

F1-micro: 0.717

F1-macro: 0.720

- **Результаты**

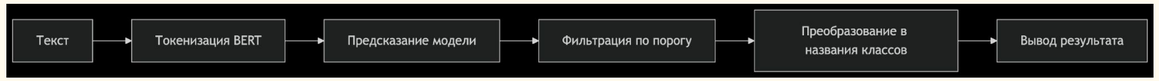
Потери (loss) снизились с 0.36 до 0.18

Модель корректно определяет классы для большинства текстов

Проблема: ложные отрицательные срабатывания — исправлено на этапе инференса модели

```
Train loss 0.17839611819497844
F1-micro: 0.7173, F1-macro: 0.7199
Validation F1 micro: 0.7172573189522342, F1 macro: 0.7199361949002757
Текст: Решили прогуляться с семьей по торговому центру. Обожаю такие семейные посиделки и все такое. Когда проголо
дались, вспомнили про Вкусс-Вилла. В нем всегда все свежее и полезное
Текст не относится ни к одному из классов
```

Приложение для классификации текстов



Назначение
Классификация текстов по 6 категориям с возможностью отнесения к нескольким классам одновременно

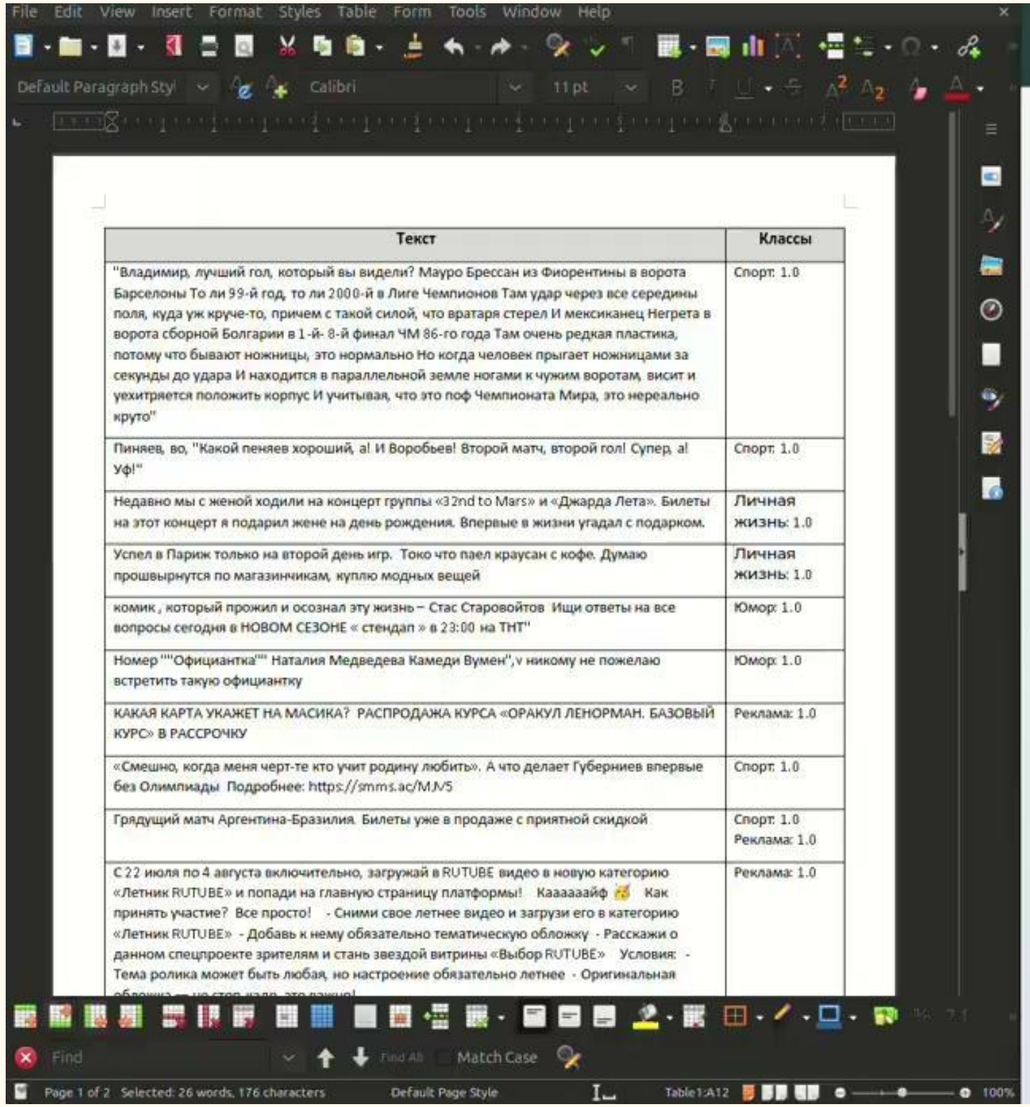
Ключевые компоненты
Модель: RuBERT (дообученный для multi-label)
Порог вероятности: 0.35
Вход: Текст до 512 токенов
Выход: Названия классов (например, "Политика, Юмор")

Как работает
Текст → токенизация BERT
Модель предсказывает вероятности для каждого класса
Фильтрация по порогу → итоговые метки

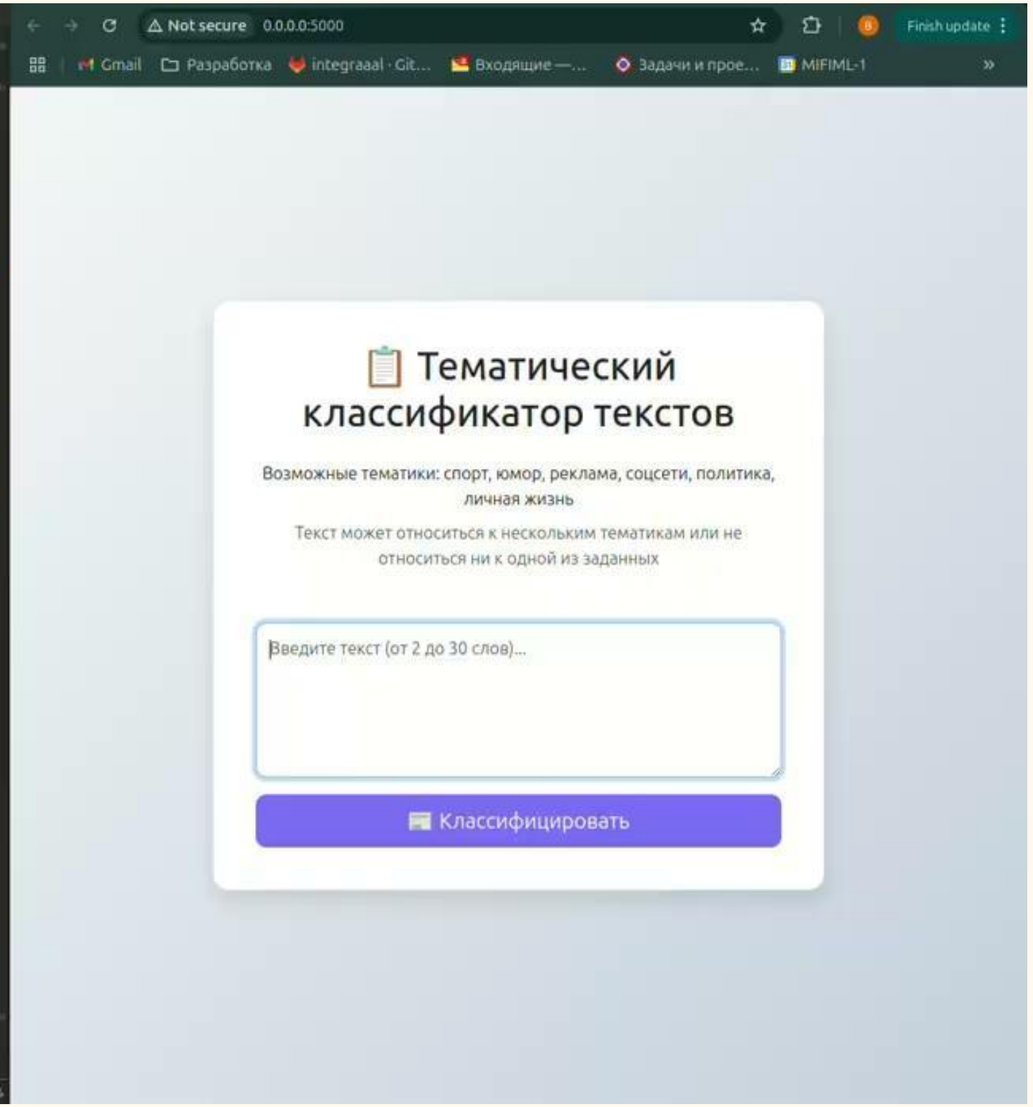
Пример
Текст: "Смешной мем про выборы! 😂"
Прогноз: "Политика, Юмор"

```
30        # Токенизация текста
31        encoding = tokenizer(
32            [text],
33            max_length=max_len,
34            truncation=True,
35            padding=True,
36            return_tensors='pt'
37        ).to(device)
38
39        # Получение предсказаний
40        with torch.no_grad():
41            outputs = model(**encoding)
42            logits = outputs.logits
43            probs = torch.sigmoid(logits)
44            predictions = (probs >= threshold).cpu().numpy()
45        print(probs)
46        # Преобразование в индексы классов
47        pred_indices = np.where(predictions[0])[0]
48
49        # Преобразуем numpy int64 в обычные питоновские int и затем в имена классов
50        if len(pred_indices) > 0:
51            # Преобразуем numpy.int64 в стандартные питоновские int
52            pred_indices = [int(idx) for idx in pred_indices]
53
54            # Добавим отладочную информацию
55            print(f"Предсказанные индексы: {pred_indices}")
56            # Проверка индексов и получение имен классов
57            try:
58                pred_classes = [int(mlb.classes_[idx]) for idx in pred_indices]
59                mapped_classes = [CLASS_MAPPING[idx] for idx in pred_classes]
60                if not mapped_classes:
61                    result = "Класс не определён"
62                else:
63                    result = ', '.join(mapped_classes)
64
65                return result
66            except IndexError as e:
67                print(f"Ошибка при доступе к классам: {e}")
68                # В случае ошибки возвращаем индексы (для отладки)
69                return [f"Класс #{idx}" for idx in pred_indices]
70
71        return "Класс не определён"
```

Демонстрация проекта



Текст	Классы
"Владимир, лучший гол, который вы видели? Мауро Брессан из Фиорентины в ворота Барселоны То ли 99-й год, то ли 2000-й в Лиге Чемпионов Там удар через все середины поля, куда уж круче-то, причем с такой силой, что вратаря стерел И мексиканец Негрета в ворота сборной Болгарии в 1-й- 8-й финал ЧМ 86-го года Там очень редкая пластика, потому что бывают ножницы, это нормально Но когда человек прыгает ножницами за секунды до удара И находится в параллельной земле ногами к чужим воротам, висит и уехитряется положить корпус И учитывая, что это поф Чемпионата Мира, это нереально круто"	Спорт: 1.0
Пиняев, во, "Какой пеняев хороший, a! И Воробьев! Второй матч, второй гол! Супер, a! Уф!"	Спорт: 1.0
Недавно мы с женой ходили на концерт группы «32nd to Mars» и «Джарда Лета». Билеты на этот концерт я подарил жене на день рождения. Впервые в жизни угадал с подарком.	Личная жизнь: 1.0
Успел в Париж только на второй день игр. Токо что паел краусан с кофе. Думаю прошвырнутся по магазинчикам, куплю модных вещей	Личная жизнь: 1.0
комик , который прожил и осознал эту жизнь – Стас Старовойтов Ищи ответы на все вопросы сегодня в НОВОМ СЕЗОНЕ « стендап » в 23:00 на ТНТ"	Юмор: 1.0
Номер ""Официантка"" Наталия Медведева Камеди Вумен",v никому не пожелаю встретить такую официантку	Юмор: 1.0
КАКАЯ КАРТА УКАЖЕТ НА МАСИКА? РАСПРОДАЖА КУРСА «ОРАКУЛ ЛЕНОРМАН. БАЗОВЫЙ КУРС» В РАССРОЧКУ	Реклама: 1.0
«Смешно, когда меня черт-те кто учит родину любить». А что делает Губерниев впервые без Олимпиады Подробнее: https://smms.ac/MJV5	Спорт: 1.0
Грядущий матч Аргентина-Бразилия. Билеты уже в продаже с приятной скидкой	Спорт: 1.0 Реклама: 1.0
С 22 июля по 4 августа включительно, загружай в RUTUBE видео в новую категорию «Летник RUTUBE» и попади на главную страницу платформы! Кааааайф 🥳 Как принять участие? Все просто! - Сними свое летнее видео и загрузи его в категорию «Летник RUTUBE» - Добавь к нему обязательно тематическую обложку - Расскажи о данном спецпроекте зрителям и стань звездой витрины «Выбор RUTUBE» Условия: - Тема ролика может быть любая, но настроение обязательно летнее - Оригинальная	Реклама: 1.0



Мы на GitHub: https://github.com/martetten/Dataton_2

Вывод

- Разработано приложение приложение с использованием предобученной модели машинного обучения, которое даёт возможность пользователю определить тематику введённого текста: спорт, юмор, реклама, соцсети, политика, личная жизнь

- Возможности развития проекта:

- Улучшение модели классификации
  - Увеличение датасета, расширение спектра категорий
  - Использование альтернативных моделей вместо DeepPavlov/rubert-base-cased для повышения точности. Например, DeepPavlov/rubert-large-cased, ai-forever/ruRoberta-large
  - Дообучение модели на специфичных данных, создание размеченного датасета
- Улучшение UI/UX
  - Сохранение истории запросов и добавление возможности экспорта результатов — CSV или JSON
  - Визуализация уверенности модели — графики, heatmap по темам

Спасибо за внимание!

Команда хакатона

SKILLFACTORY