Промежуточная защита

Команда 28

Тема: Предсказание движения цен на фьючерсы на основе текстовых данных

Руководитель: Ковалева Александра
Члены команды: Рябцев Василий, Фазилов Сергей, Хоменко Павел, Константинов Артем

Постановка задачи

Необходимо разработать модель для предсказания динамики изменения цен акций на основе различных текстовых данных.

К текстовым данным относятся: посты из реддита и твиттера, заголовки новостей, отчеты компаний по форме 10-k.

Также для модели необходимо разработать приложение с архитектурой клиент – сервер, на основе фреймворков FastAPI и Streamlit.

Цели на год

Сбор и обработка данных:

- Отбор компаний и источников текстовых данных;
- Парсинг и разметка данных;

Машинное обучение:

- Разведывательный анализ данных (EDA);
- Выбор алгоритмов и метрик;
- Обучение моделей и отбор наилучшего решения.

Глубокое обучение:

- Разработка архитектуры нейросети;
- Настройка и обучение модели;
- Оценка и улучшение качества модели.

Приложение:

- FastAPI: Разработка API для взаимодействия с моделью (эндпоинты, обработка данных, возврат результатов).
- Streamlit: Создание интерфейса для загрузки данных и визуализации результатов модели.
- Деплой: Развертывание приложения в Docker, тестирование.

Прогнозирование цены акции по твитам инфлюенсеров

Общая информация о датасете

- 527 твитов;
- 11 атрибутов;
- твиты 71 инфлюенсера из списков forbes.com, thecfoclub.com, hypefury.com;
- твиты за последние 10 лет по 63 компаниям из разных секторов экономики;
- скрапинг и парсинг с помощью библиотеки twikit;
- английский язык

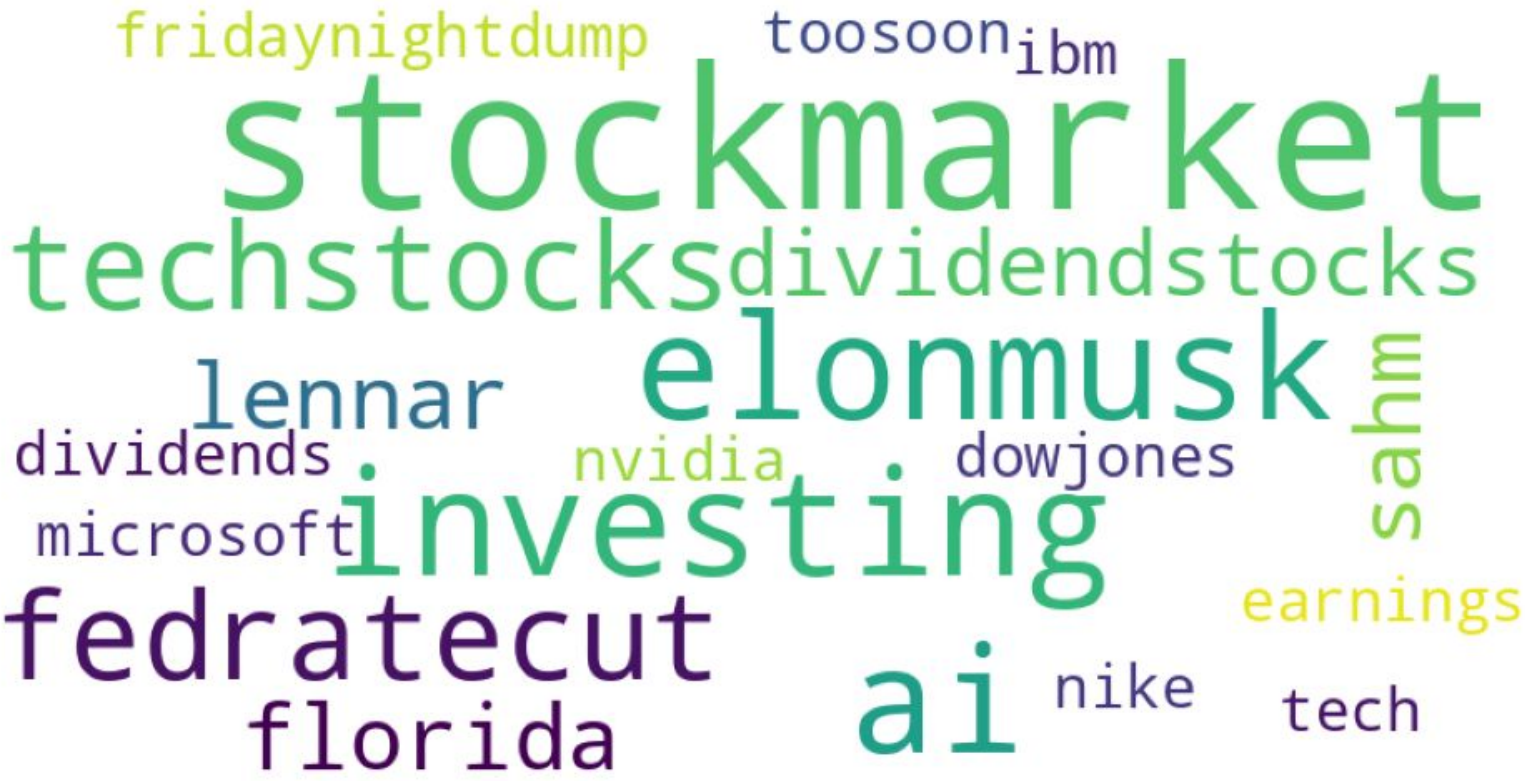
Описание атрибутов

1. **text** – текст твита, заголовок превью от ссылки (если есть), список хэштегов (если есть). str
2. **is_quote_status** – указывает на наличие у твита статуса цитаты. bool
3. **view_count** – кол-во просмотров (пропуски заполнялись медианой). float
4. **has_card** – указывает, содержит ли твит карточку. bool
5. **urls** – указывает, содержит ли твит ссылку. bool
6. **day** – день публикации. int
7. **month** – месяц публикации. int
8. **year** – год публикации. int
9. **is_in_reply_to** – указывает, является ли твит ответов на другой твит. bool
10. **is_view_count** – указывает, доступна ли информация о просмотрах твита. bool
11. **1_day_after** – таргет: 1 - на следующей день цена тикера, упомянутого в твите, выросла, 0 - на следующий день цена тикера, упомянутого в твите, снизилась. int

Самые популярные слова в твитах

Самые популярные слова из превью ссылок

Самые популярные хэштеги



Количество просмотров

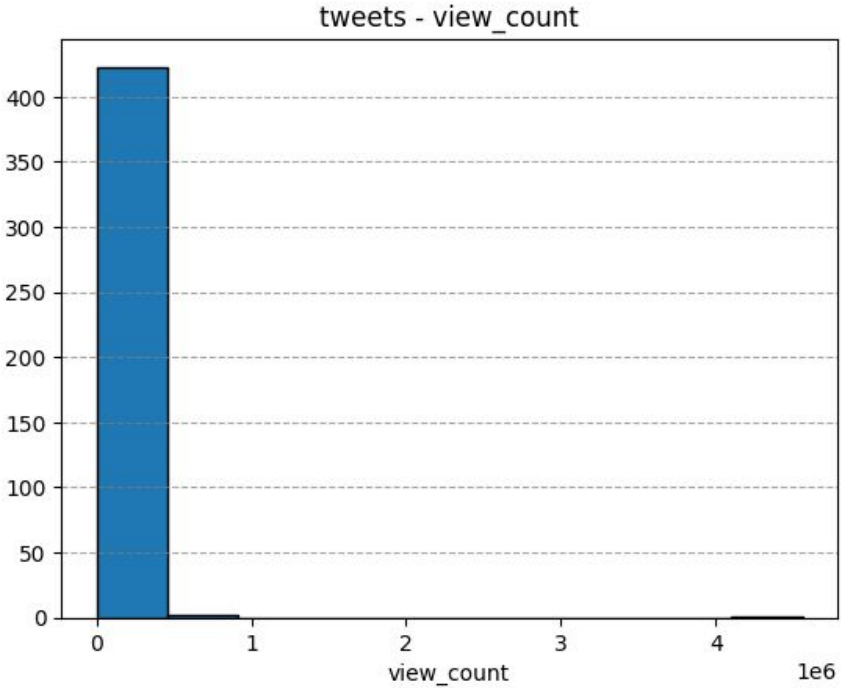


Рис. 1 – Твиты - кол-во просмотров

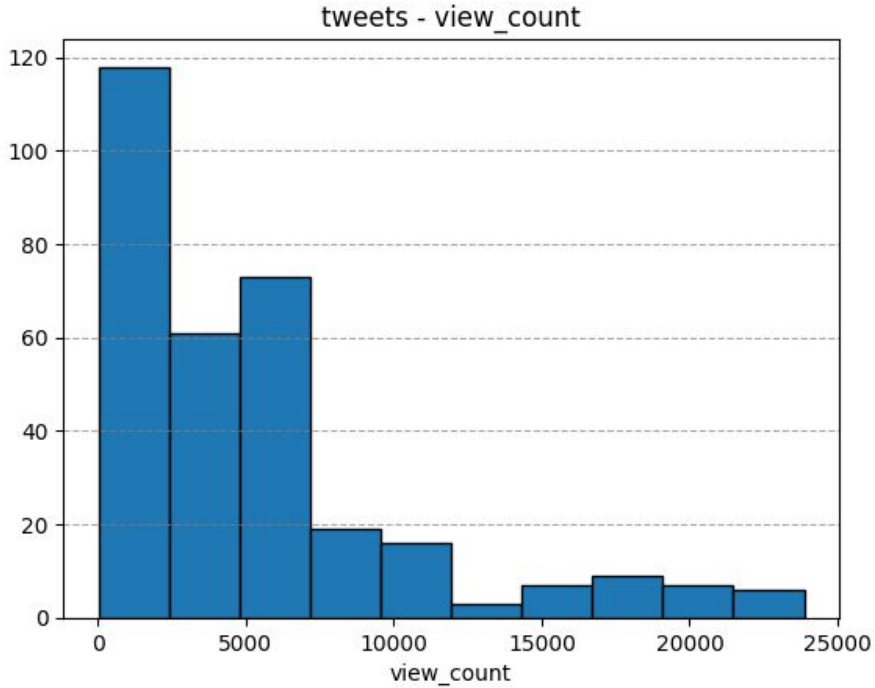
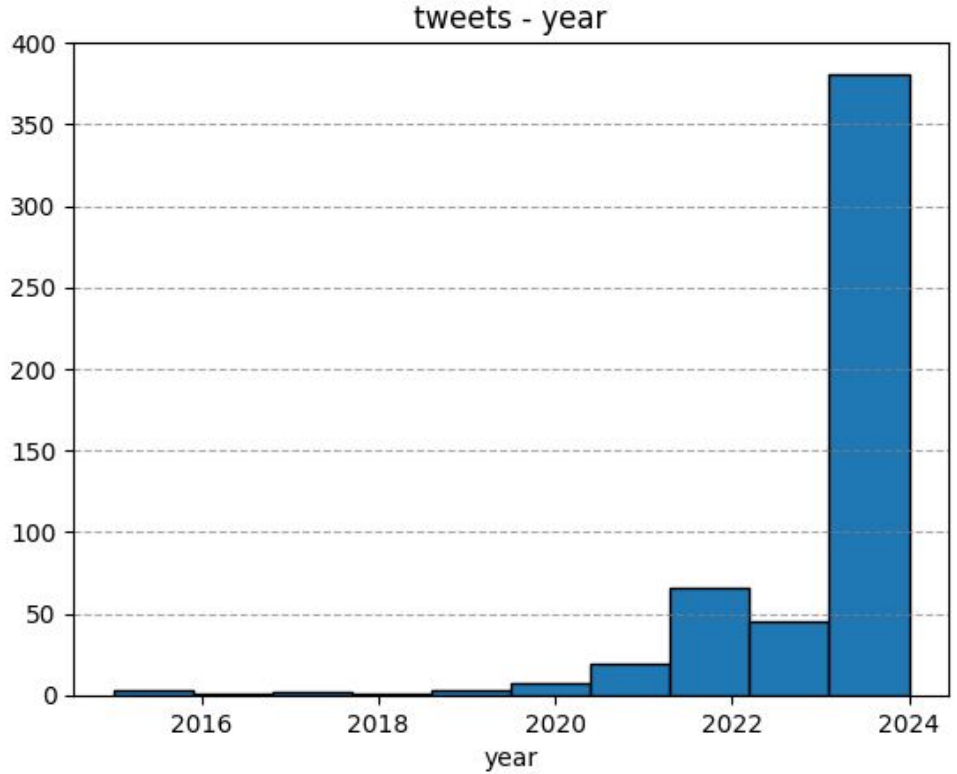


Рис. 2 – Твиты - кол-во просмотров, 0.75 квантиль

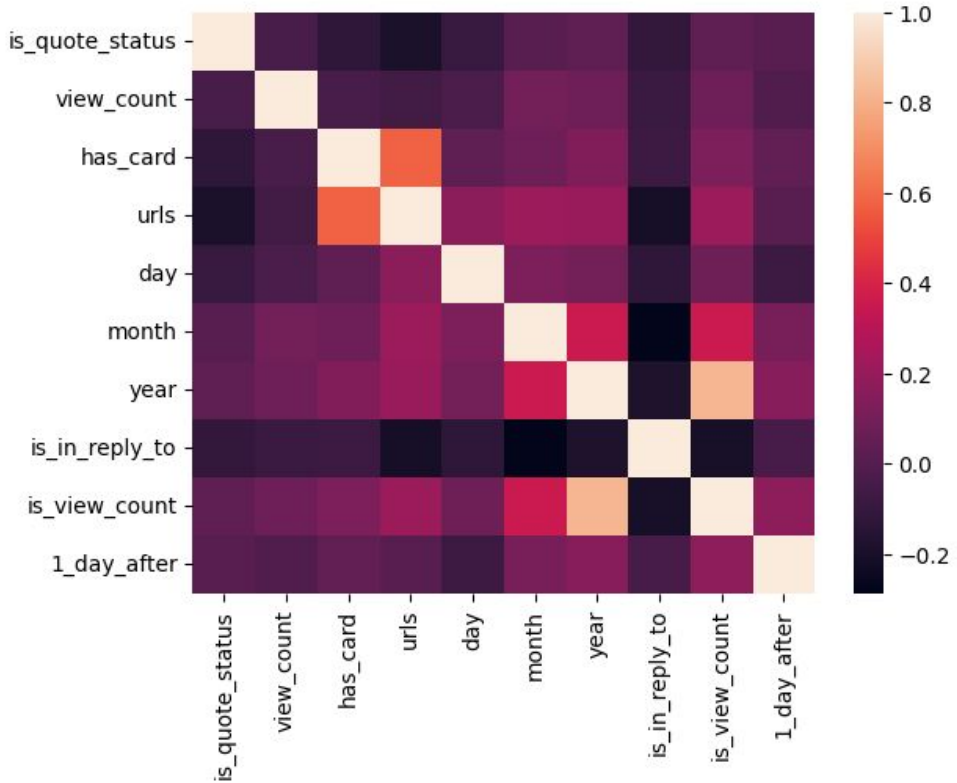
Количество твитов по годам



Соотношение классов таргета

target class	proportion
1	0.53
0	0.47

Корреляция Пирсона



Baseline решение

Тип задачи: двухклассовая классификация

Метрики: ROC-AUC

Семейство моделей: Логистическая регрессия

Векторизаторы текстов: BoW, Tf-idf, Предобученный ProsusAI, Word2vec

Нормировка: StandardScaler

Доля теста: 0.1

Кол-во фолдов для кля кросс валидации: 5

Подбор гиперпараметров через GridSearchCV

BoW: max_features, ngram_range

Tf-idf: max_features, ngram_range, smooth_idf

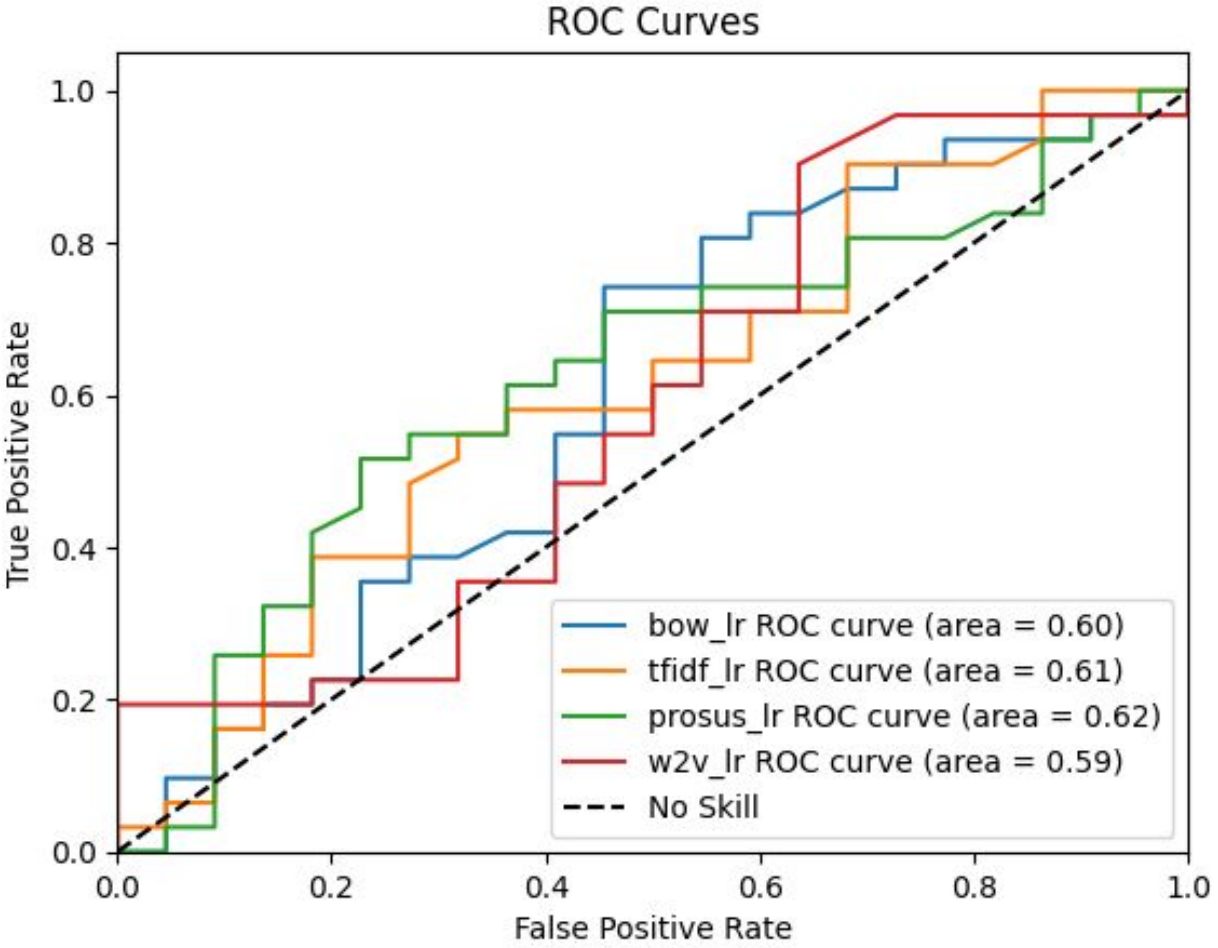
LogisticRegression: C, penalty, solver, max_iter

Результаты обучения моделей на текстовых признаках

model name	ROC-AUC train	ROC-AUC test	std
bow_lr	0.99	0.60	0.06
tfidf_lr	0.96	0.61	0.03
prosus_lr	0.57	0.62	0.06
w2v_lr	0.64	0.59	0.08

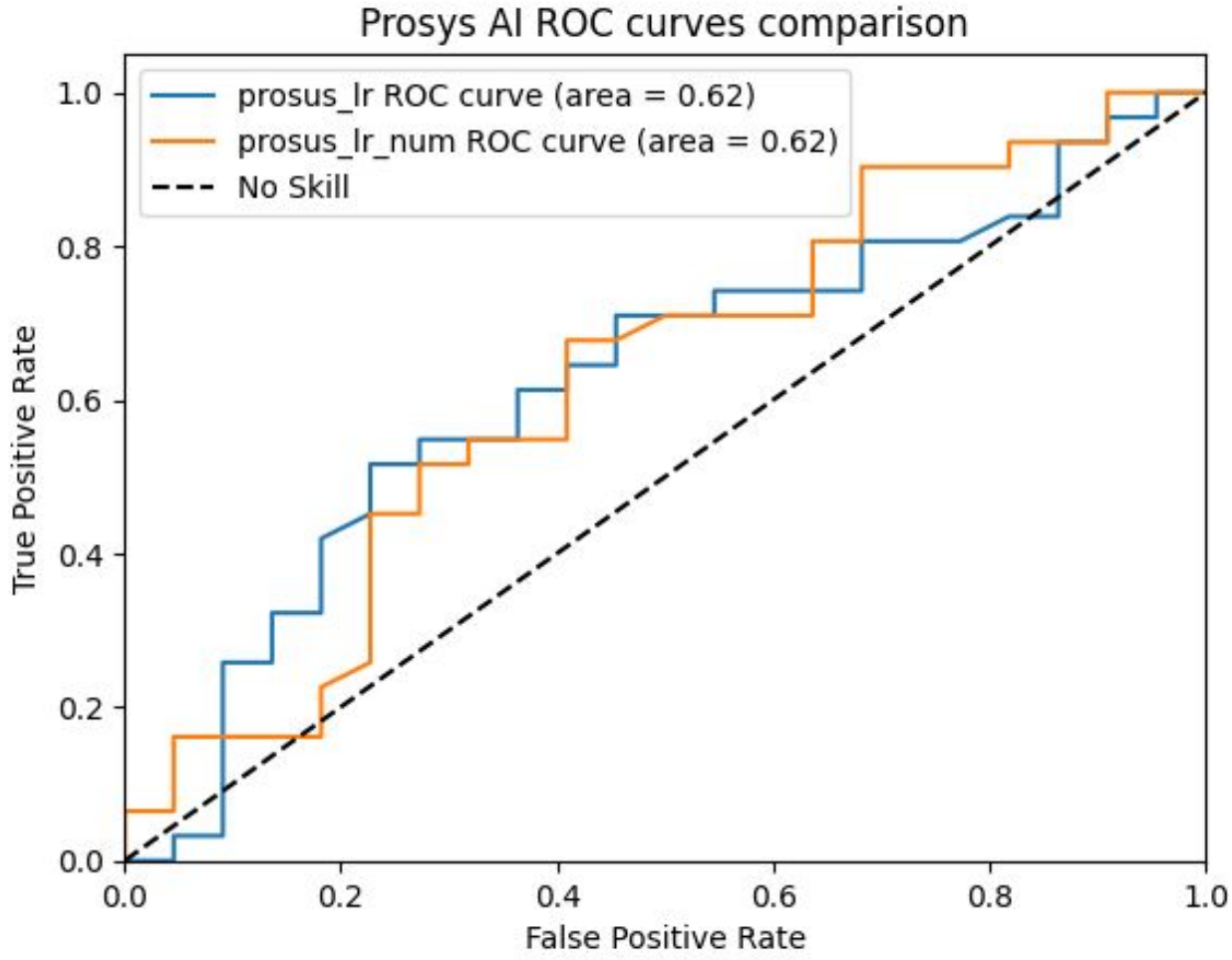
model name: <векторизатор>_<алгоритм> [_<версия>]

std рассчитывалось для оценок по кросс валидации для моделей с подобранными параметрами


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
bow_lr ROC curve (area = 0.60)
tfidf_lr ROC curve (area = 0.61)
prosus_lr ROC curve (area = 0.62)
w2v_lr ROC curve (area = 0.59)
No Skill

Сравнение prosus_lr обученной на текстовых признаках и на всех признаках

Model name	ROC-AUC train	ROC-AUC test	std
prosus_lr	0.57	0.62	0.06
prosus_lr_num	0.62	0.62	0.07


Prosys AI ROC curves comparison
prosus_lr ROC curve (area = 0.62)
prosus_lr_num ROC curve (area = 0.62)
No Skill
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Планы на второе полугодие

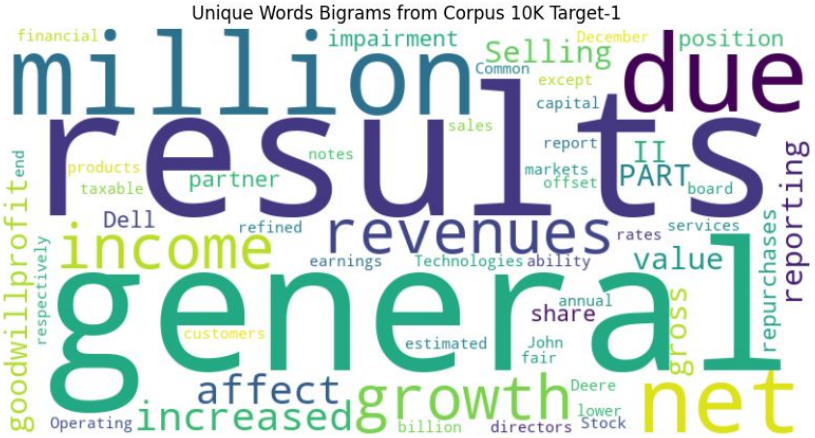
1. Собрать больше данных
2. Feature engineering
3. DL методы
4. Интерпретация предсказаний модели

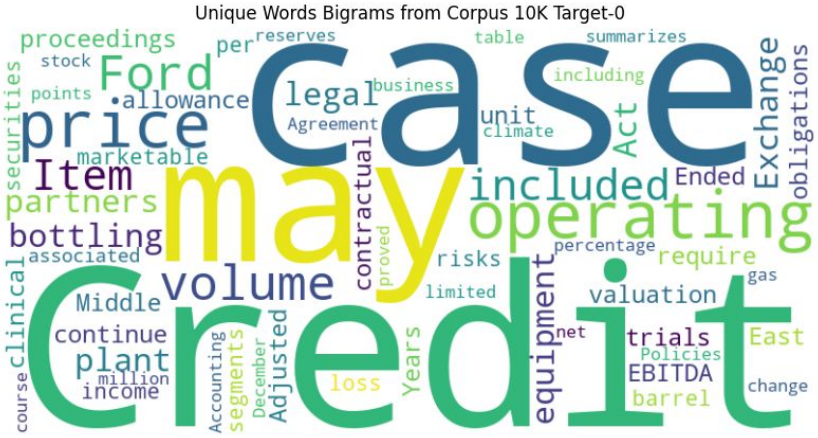
Прогнозирование цены акции по корпоративным отчетам

Исходные данные для исследования

- 1 476 отчетов 10-К (годовой корпоративный отчет) по 112 компаниям из разных секторов экономики за 2004-2023 годы
- Из-за большой длины отчета (в среднем 60 тыс. слов) и сложности машинной обработки большого датасета (2,2 Гб) для построения модели был использован только раздел “Management Discussion and Analysis (MD&A)
- MD&A содержит анализ финансовых результатов компании, планы и риски, которые могут повлиять на будущее
- Поскольку большинство трансформеров может обрабатывать только 512 токенов, методами оптимизации был определен участок MD&A, дающий наибольший Accuracy

EDA





Выше приведены словы из топ-300 биграмм, которые есть в одной группе таргета, но отсутствуют в другой:

- в акциях, которые росли после публикации отчетности, встречались уникальные слова с позитивной коннотацией: results, growth, increased, revenues, income, earnings, value, profit
- у падающих компаний в биграммах встречали слова с негативным подтекстом: credit, limited, trials, obligations

Результаты моделирования (1/2)

Модель	Target	Фрагмент текста	Accuracy (train)	Accuracy (test)
BoW (unigrams)	target_3	22500:24500	100%	59%
BoW (bigrams)	target_3	45500:48500	100%	60%
BoW (trigrams)	target_3	56500:59500	88%	59%
TF-IDF (unigrams)	target_3	500:3500	79%	59%
TF-IDF (bigrams)	target_3	60500:66500	87%	56%

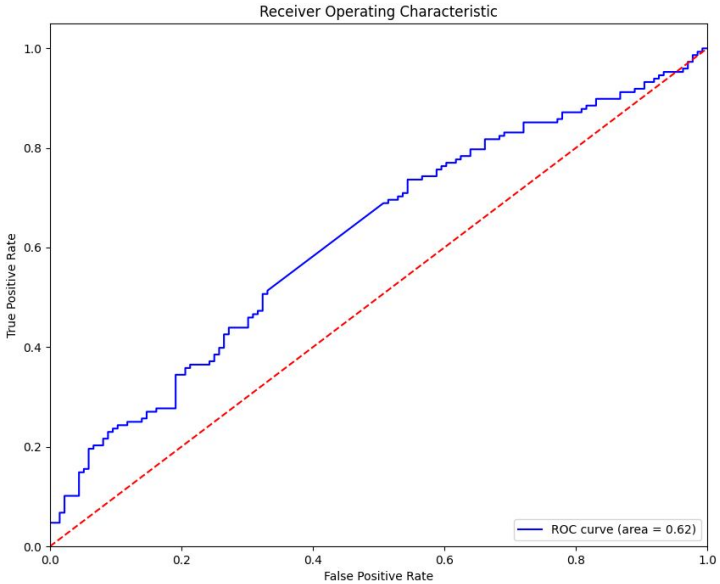
Результаты моделирования (2/2)

Модель	Target	Фрагмент текста	Accuracy (train)	Accuracy (test)
TF-IDF (trigrams)	target_3	22500:24500	80%	59%
BERT (DistilBERT)	target_10_index	45500:48500	70%	60%
BERT (FinBERT)	target_10_index	56500:59500	74%	59%
GPT-2	target_1_index	500:3500	74%	59%

Итоговая модель

В качестве итоговой модели выбран трансформер DistilBERT, так как он обладает наименьшим уровнем переобучения и является “легким” трансформером

ROC-AUC baseline модели: 0.62



Направления дальнейших исследований

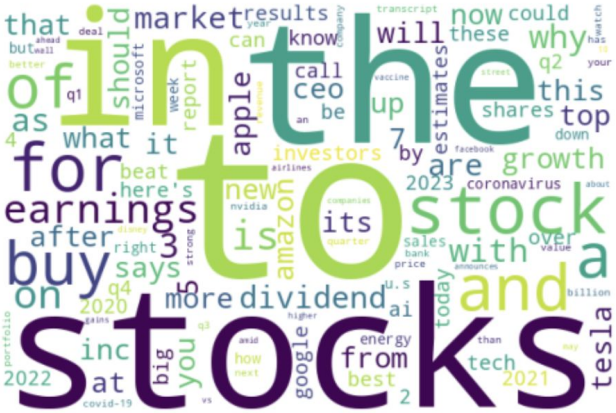
- Обогащение датасета числовыми данными - показателями финансовой отчетности, обучение модели с учетом показателей отчета
- Оптимизация гиперпараметров модели
- Обучение трансформера на данных отчетов 10K

Прогнозирование цены акции по новостям

Общая информация о датасете

- 302 объекта (тикера);
- 582800 наблюдений;
- 5 атрибутов;
- новости с более чем 15 источников: Reuters, Yahoo Finance, Forbes, Bloomberg, NY Times & etc.;
- новости за период январь 2020 - апрель 2024;
- английский язык

EDA



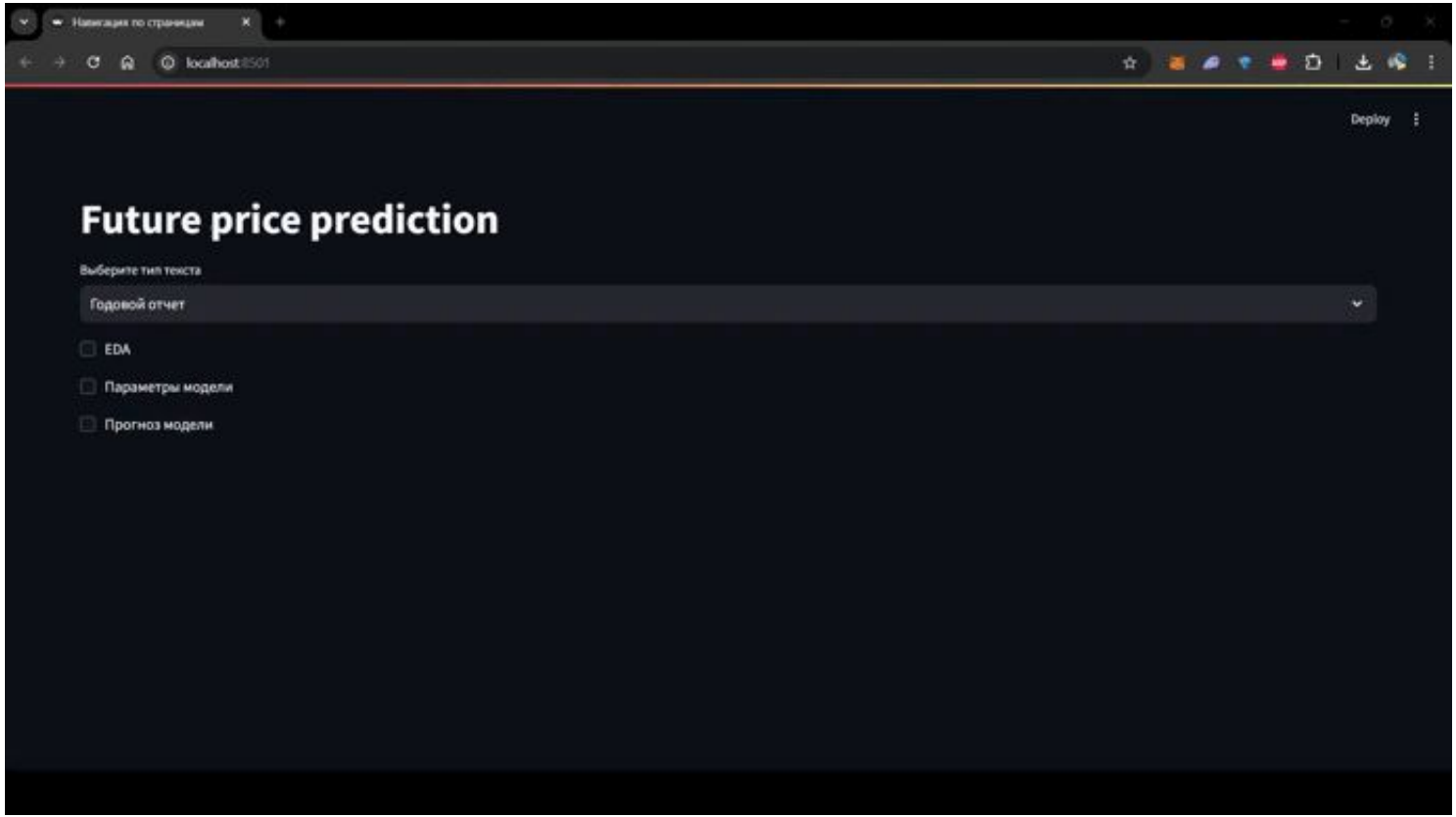
titles

text

- в акциях, которые росли после публикации новостей, в заголовках встречались уникальные слова: growth, more, earnings, best, new, apple
- в текстах аналогичные слова: result, revenue, first, strong, up, announced

Приложение Streamlit

Пример работы клиента (10k)



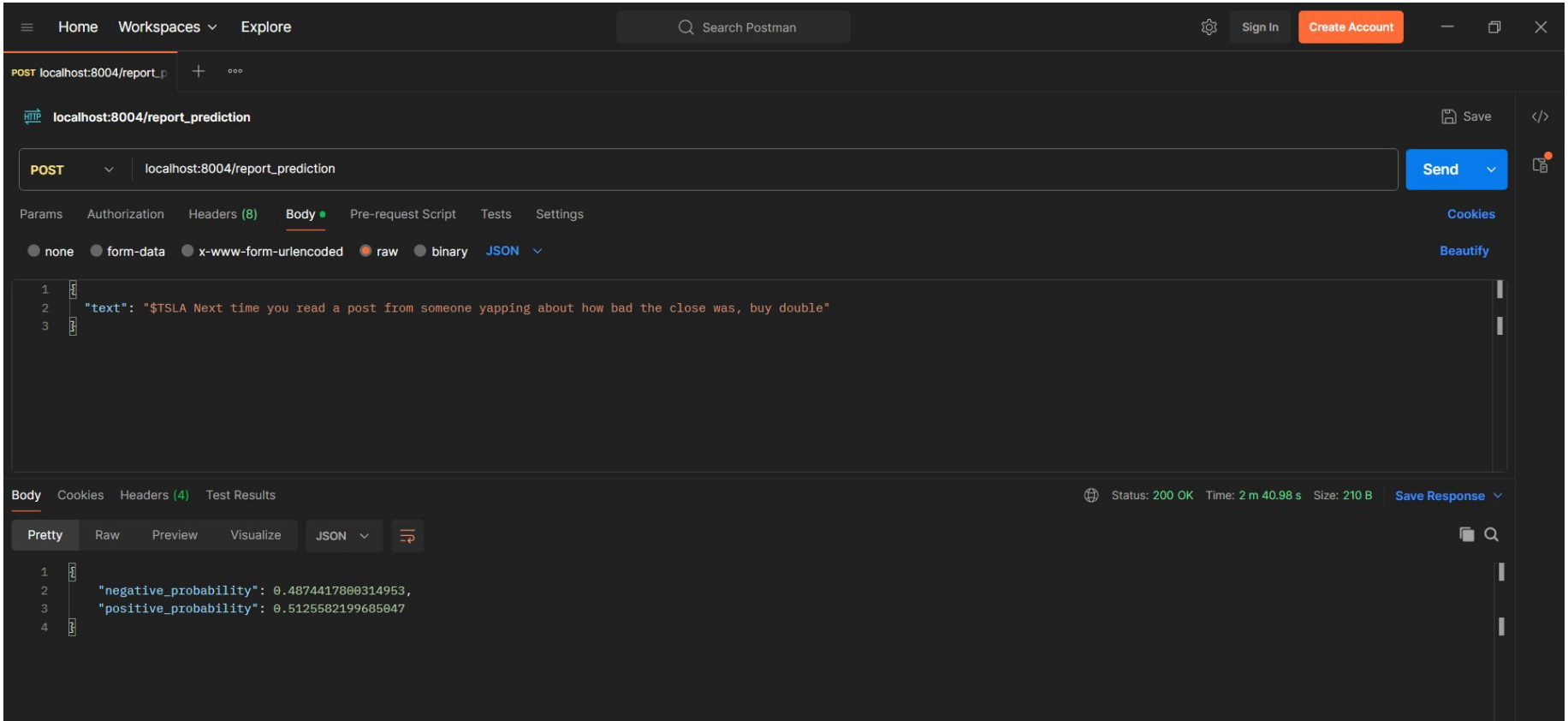
Сервисная часть на FastAPI

Как реализована сервисная часть на FastAPI

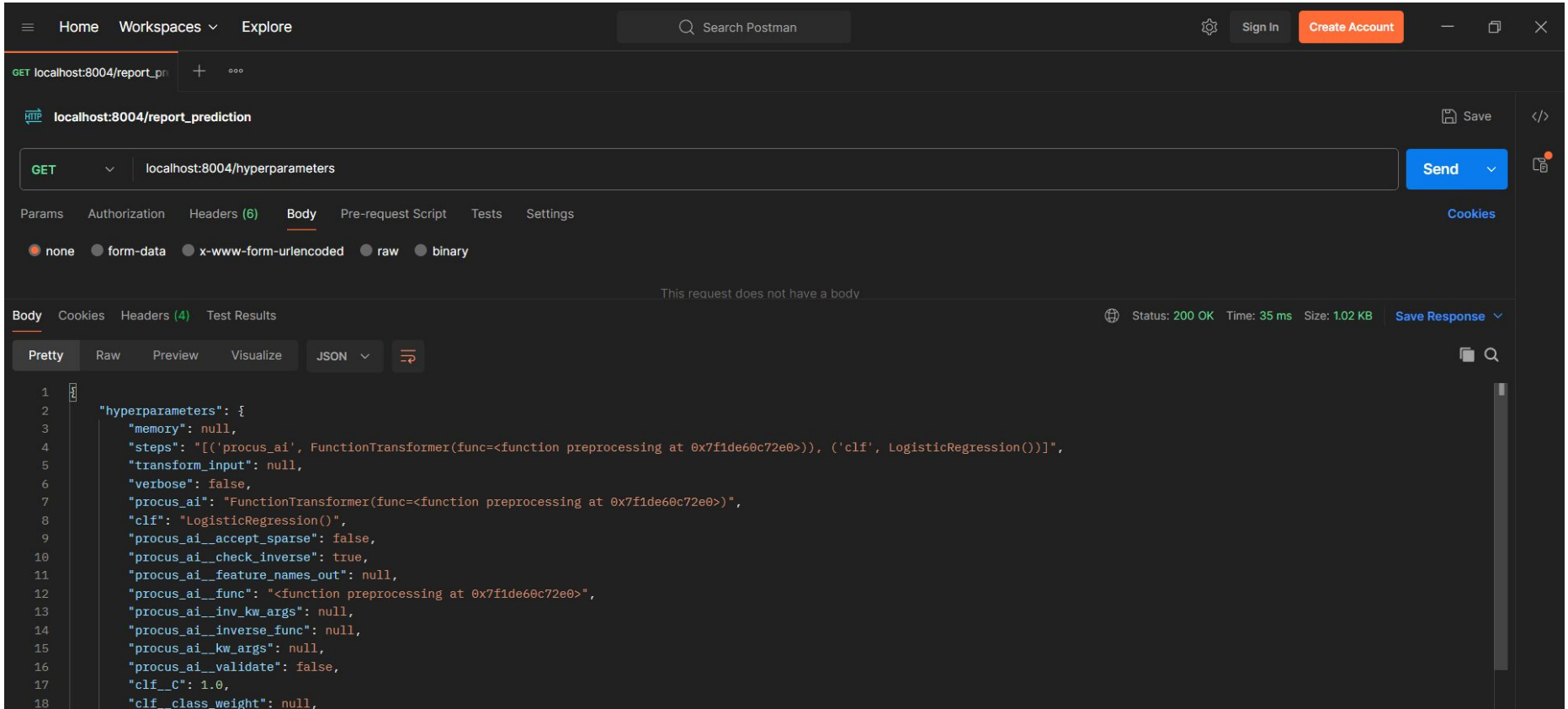
Структура кода:

1) Загрузка модели: pickle.load()
2) Обработка данных и предсказания: pd.DataFrame(), model.predict()
3) Возврат гиперпараметров: model.get_params()
4) Запуск сервера: uvicorn.run()

Демонстрация работы сервиса на FastAPI



Демонстрация работы сервиса на FastAPI



Распределение работы в команде

Рябцев Василий: парсинг и скрапинг твиттера, EDA твиттер, обучение моделей твиттер, деплой (развертывание приложения в Docker, настройка логов), разработка в FastAPI и Streamlit (частично).

Константинов Артем: парсинг и обработка датасета новостей, Streamlit.

Хоменко Павел: скрипт для получения котировок по API AlphaVantage, сбор данных и EDA датасета 10K, обучение модели, разработка FastAPI, перенос EDA в Streamlit

Фазилов Сергей: разработка FastAPI