Направление подготовки: 02.03.02
«Фундаментальная информатика и информационные технологии»

Гурин Иван Сергеевич

ОЦЕНКА ВЛИЯНИЯ НОВОСТЕЙ В СОЦИАЛЬНЫХ СЕТЯХ НА ФИНАНСОВЫЕ ПОКАЗАТЕЛИ

Научный руководитель:
Старший Преподаватель Кафедры ИВЭ - Ячменева Наталья Николаевна

Постановка задачи

Цель данной работы – создание программы, способной производить оценку влияния новостей в социальных сетях на финансовые показатели. Реализация цели разбита на следующие подзадачи:

- Поиск и создание необходимых наборов данных
- Алгоритмическая реализация применения методов предобработки к входным данным
- Применение классификатора с целью общего анализа сентимент-тональности новости
- Формирование и применение каскада регрессоров для анализа сентимент-тональности новостей
- Оценка влияния сентимент-тональности на финансовые показатели с помощью регрессивной модели
- Создание Telegram чат-бота для предоставления удобного интерфейса и гибкого взаимодействия с моделью.

Датасет FiNeS

- Более 500 новостных публикаций
- Надёжная разметка, Fleiss’ Kappa = 0,2639
- Содержит финансовые новости и сентимент-оценку к ним

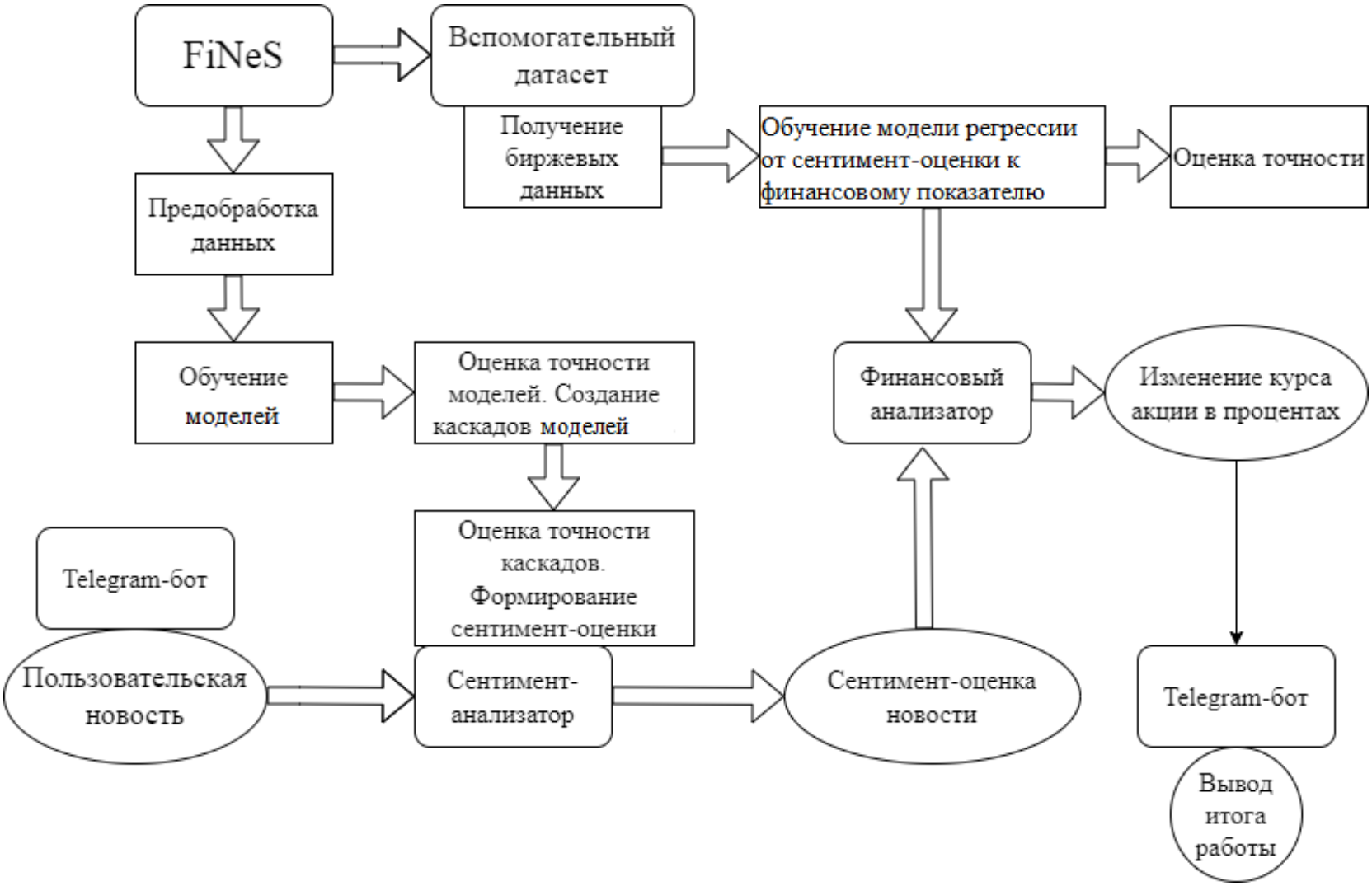
title	score
Электромобильный стартап Arrival экс-главы Yota уйдет из России	-0.5833333333
Шрёдер отклонил предложение войти в совет директоров «Газпрома»	-0.3333333333
Шельф берут в разработку // Генподрядчиком «Газпрома» на море может стать компания Андрея Патрушева	0.7
Чистая прибыль "РусГидро" по РСБУ за 1 полугодие выросла на 17%	0.6818181818
Финский производитель шин Nokian Tyres решил уйти из России	-0.4117647059
Федун ушел с поста вице-президента ЛУКОЙЛа на пенсию	-0.02564102564

Набор биржевых данных о стоимости ценных бумаг

- Содержит биржевые данные о стоимости ценных бумаг
- Отражает зависимость изменения стоимости акций от сентимент-оценкки
- Имеет 527 записей
- Имеет колонки: score, published, tickers, open, close

score	published	tickers	open	close
-0.5833333333333333	2022-05-12	ARVL	1.53	1.38
-0.3142857142857143	2022-05-20	NMTP	5.56	5.315
-0.3333333333333333	2022-05-24	GAZP	263	250.4

Этапы решения основной цели



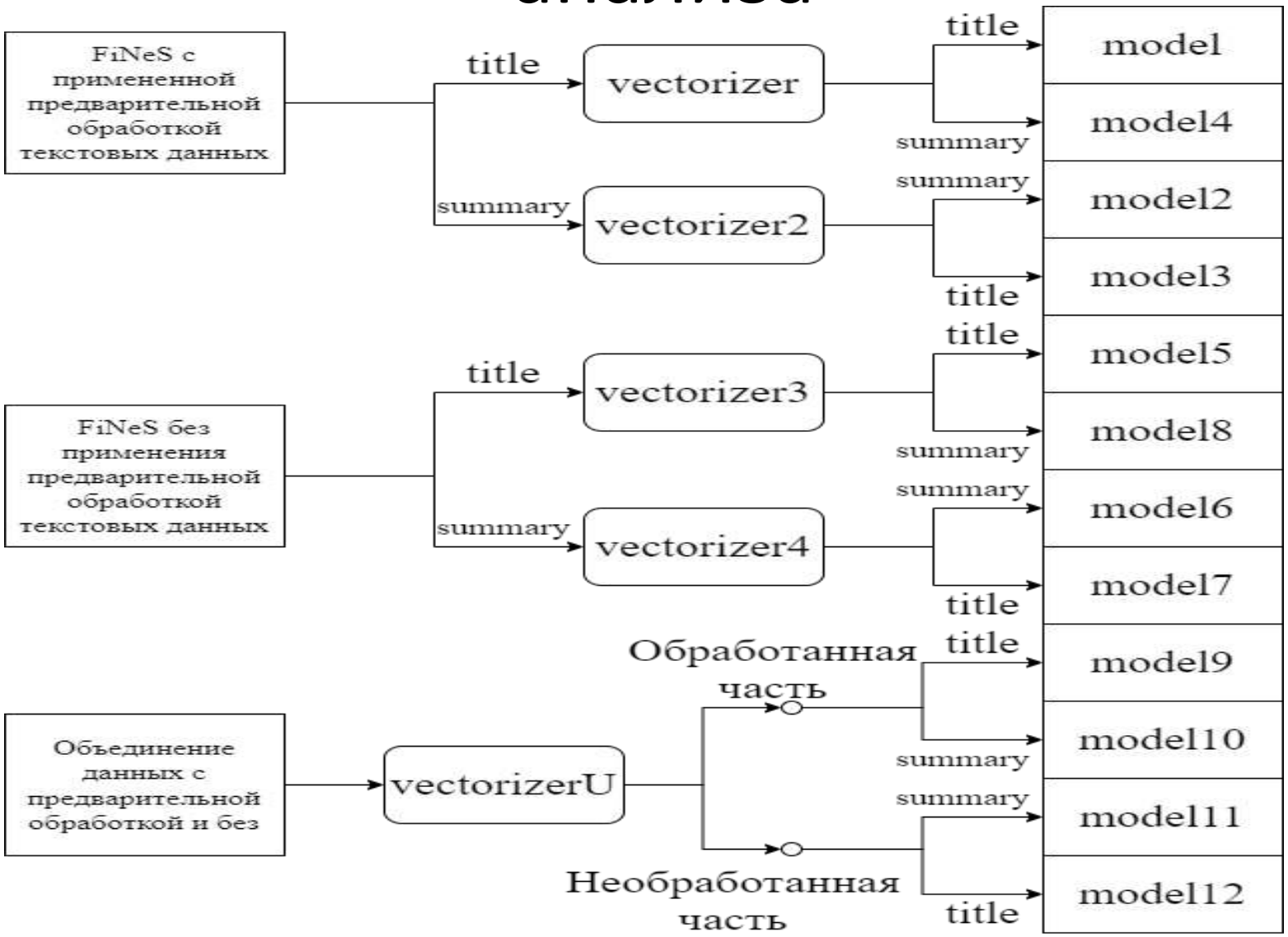
Методы предобработки данных для обучения моделей

- Заполнение пустот в FiNeS
- Токенизация по словам (nltk)
- Удаление стоп-слов (nltk)
- Лемматизация (pymorphy2)
- TF-IDF векторизация (sklearn)

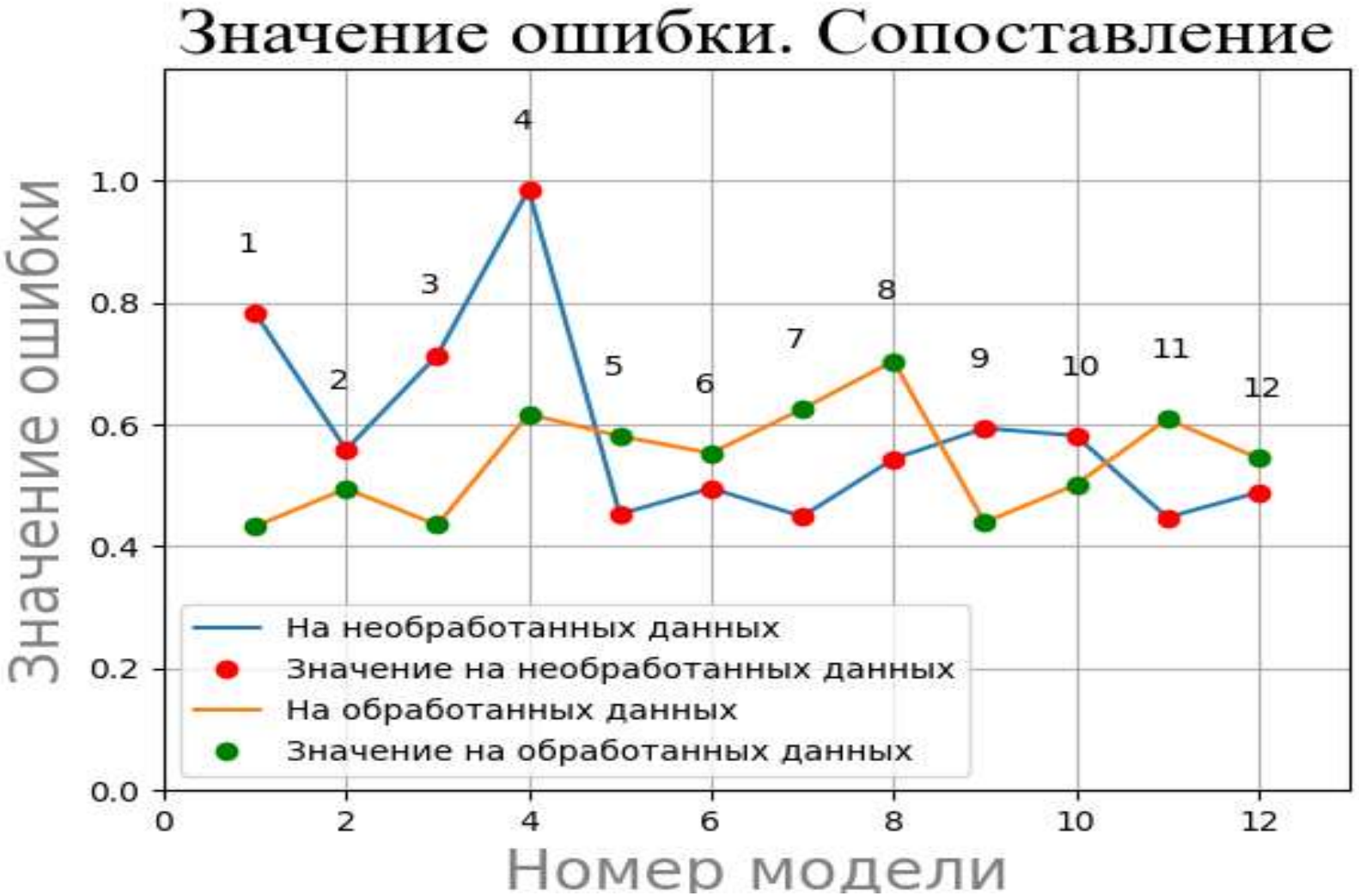
Было составлено 5 наборов данных:

- Полностью предобработанная колонка title
- Полностью предобработанная колонка summary
- Необработанная колонка title
- Необработанная колонка summary
- Объединение всех вышеперечисленных

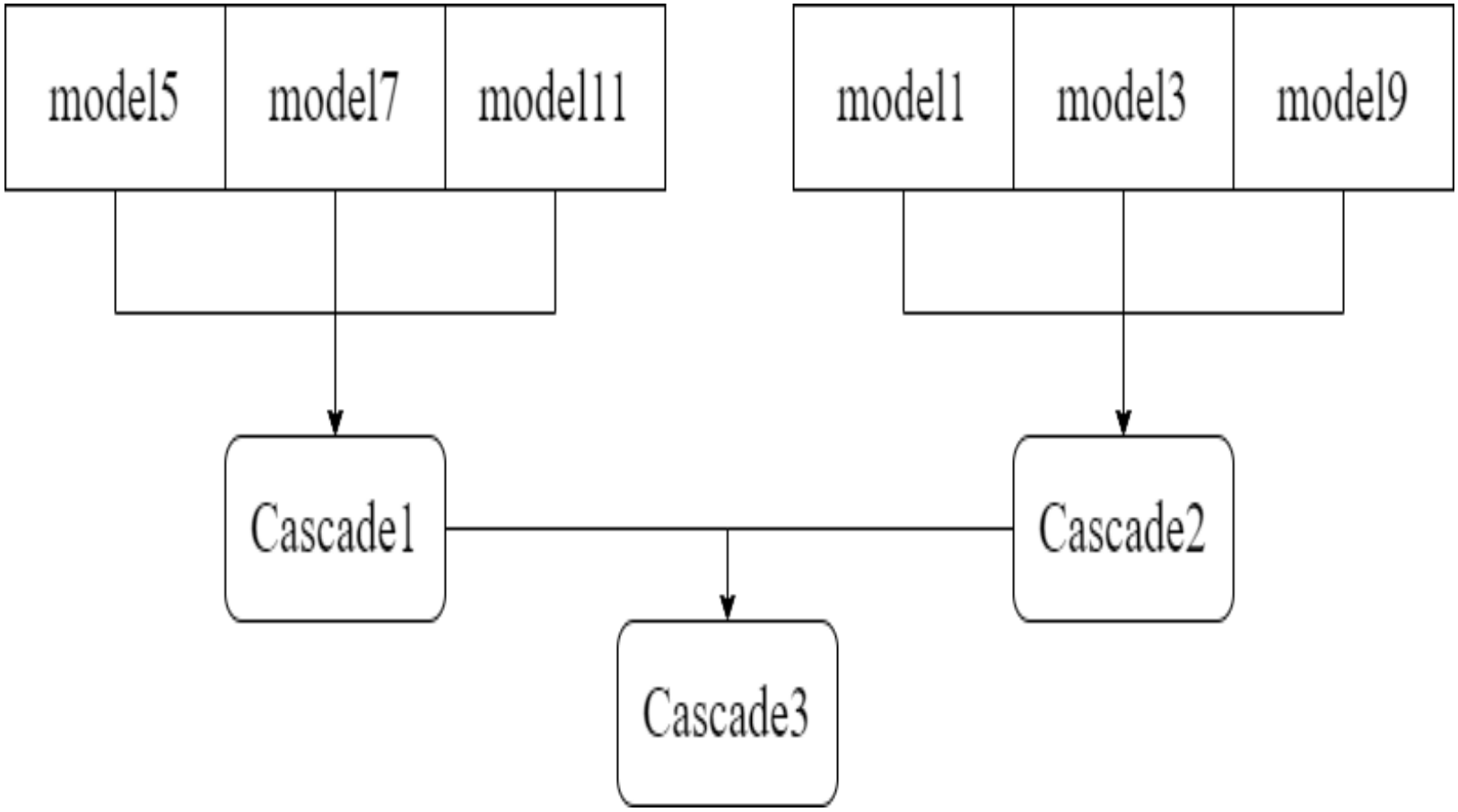
Модели регрессии для сентимент-анализа



Точность моделей регрессии

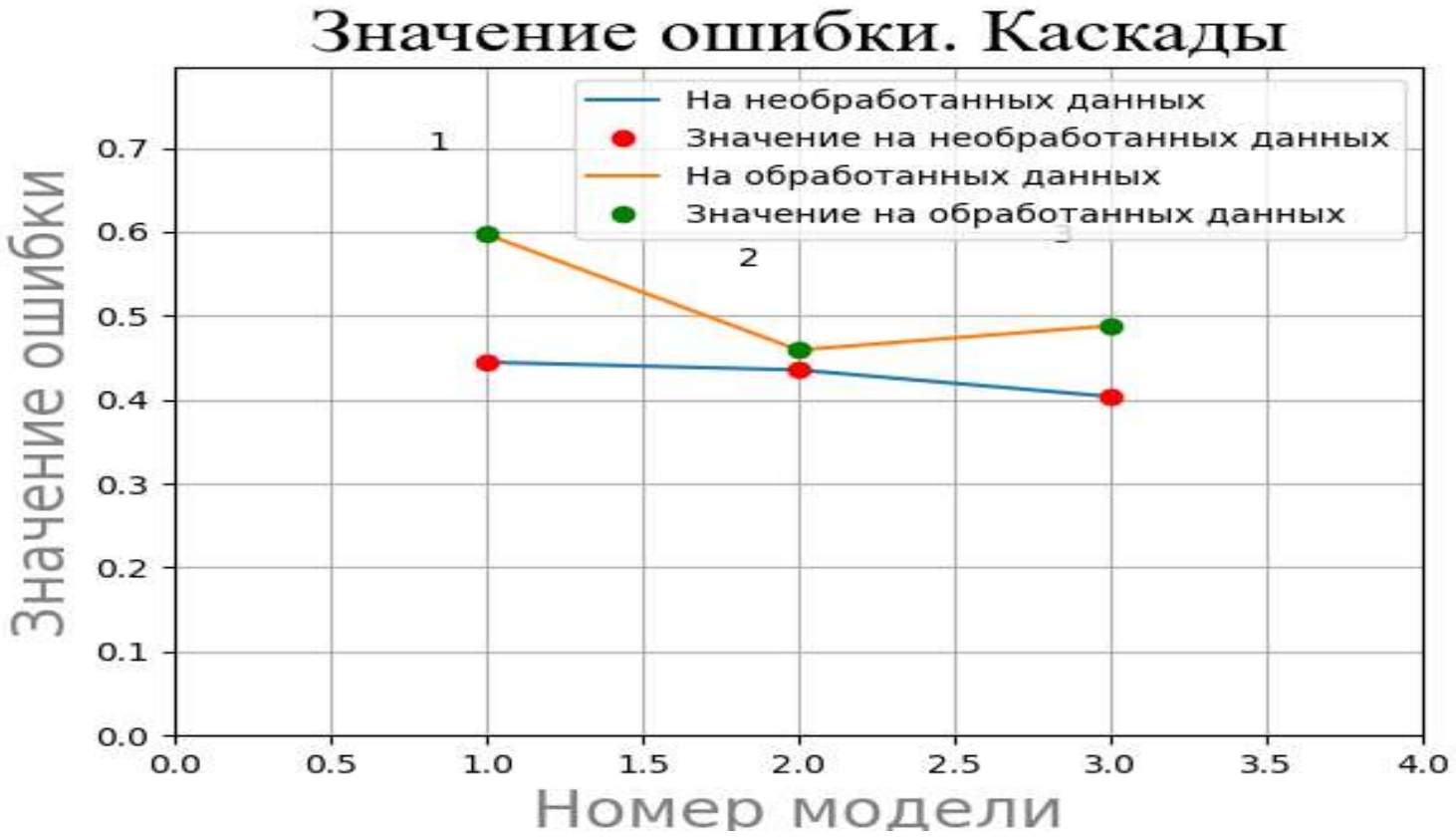


Каскады моделей



Точность каскадов моделей

Каждый каскад проверен на данных с предварительной обработкой и на данных без неё.



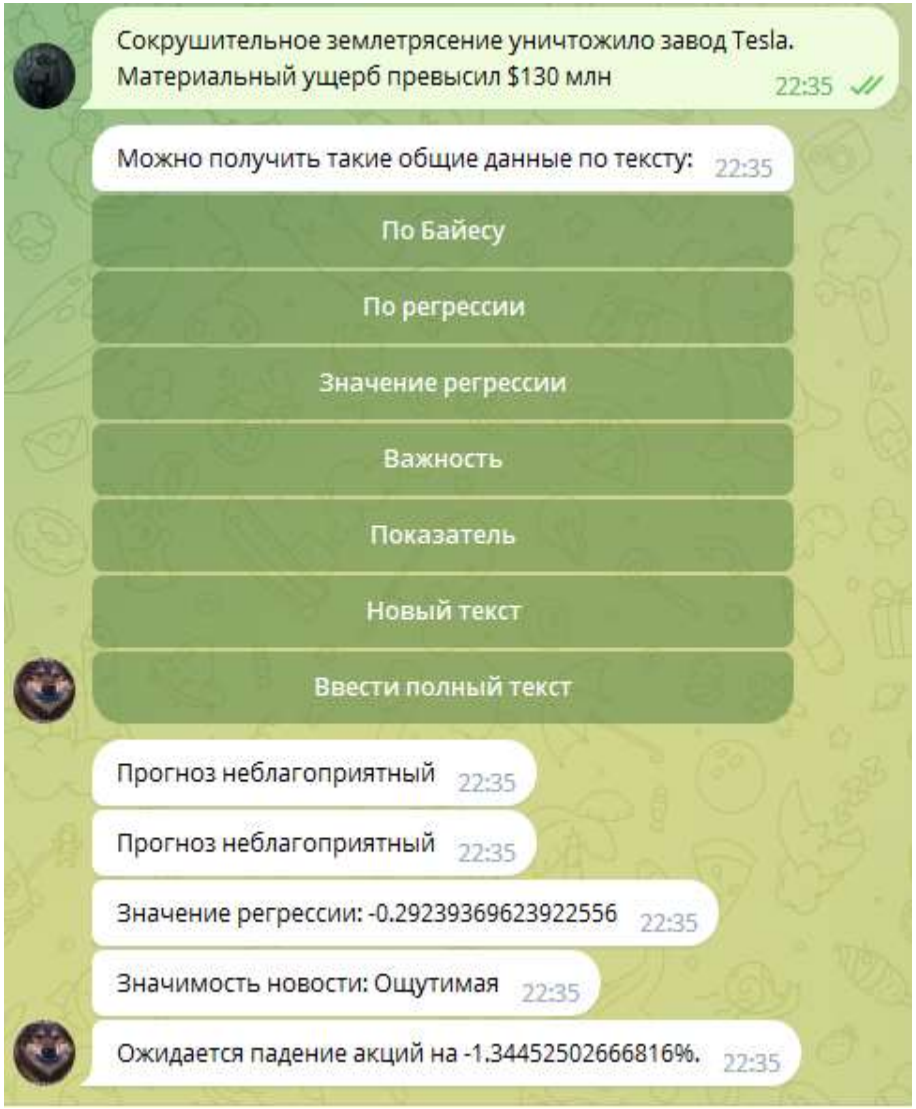
Получение итоговой оценки

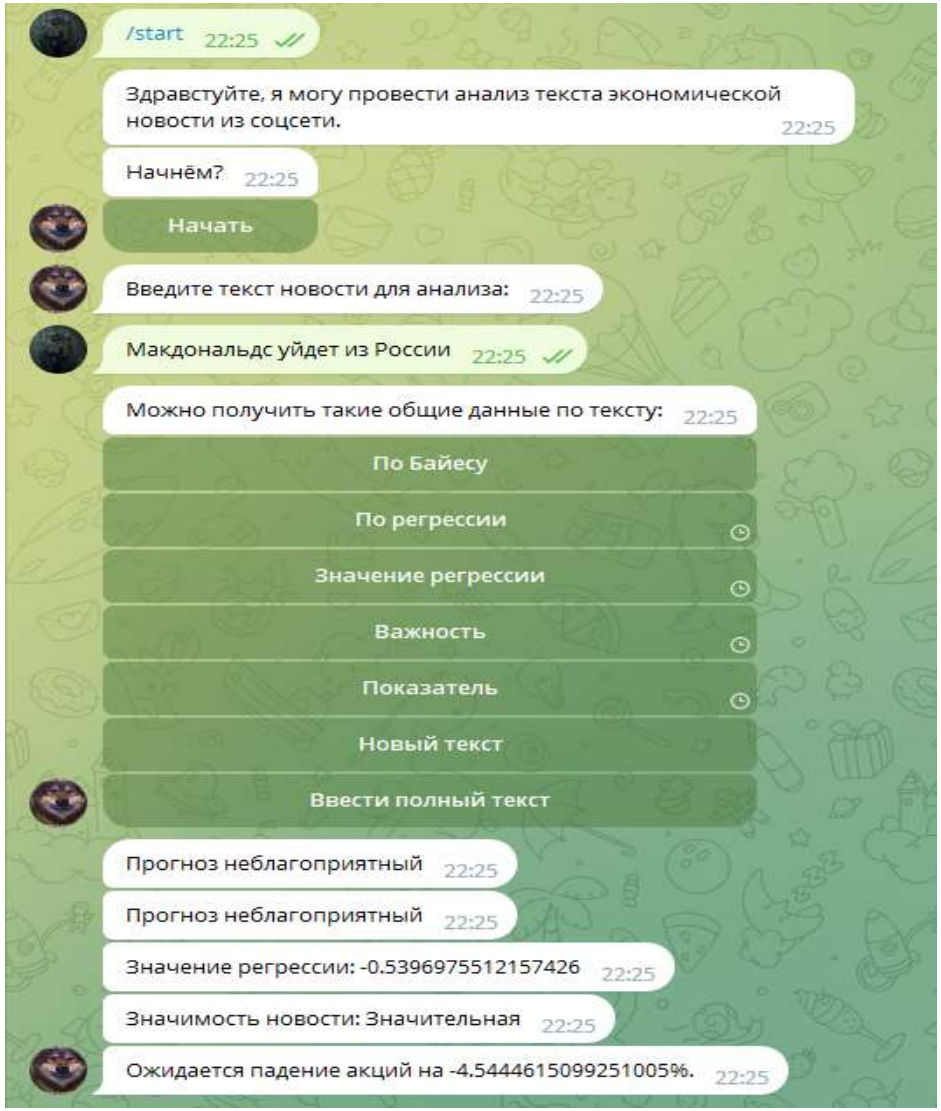
- Итоговая оценка получена в результате работы регрессионной модели, обученной на биржевых данных и данных о сентимент-оценке

Telegram-бот

- Для предоставления удобного доступа к моделям разработан Telegram-бот
- Способен принимать новостной текст и оценивать его по некоторым параметрам
- Среди параметров: оценка по Байесу, оценка по регрессии, значение регрессии, степень значимости новости, приблизительное изменение стоимости акций некоторой упоминаемой в новости компании в процентном выражении

Примеры работы приложения





Задействованные инструменты

- Nltk
- Sklearn
- Pymorpy2
- Pandas
- PyTelegramBotAPI
- FiNeS
- Python 3.7
- Ms Visual Studio

Полученные результаты

- Был реализован алгоритм, способный производить оценку влияния новостей в социальных сетях на финансовые показатели
- Были подобраны и созданы датасеты, необходимые для достижения цели
- Программно реализовано применение методов предобработки текстовых данных.
- Выбраны и применены методы классификации и анализа финансово-новостных данных.
- Сформированы, обучены и применены каскады регрессивных моделей.
- Произведена оценка влияния сентимент-тональности на финансовые показатели.
- Программно реализован Telegram-бот для удобства применения обученных моделей.

Ссылка на репозиторий