



Презентация проекта

Создание прототипа **ИИ-системы** по анализу текстовых отзывов населения с кластеризацией и визуализацией

Цель проекта

Разработка Streamlit-приложения для:



определения тональности отзывов
(позитив/негатив/нейтрал)



оценки уверенности модели



кластеризации отзывов с помощью PCA и BERT



генерации облаков слов по тональностям



Интерактивное приложение для анализа
текстовых отзывов

Используемые технологии

Технологический стек для анализа текстовых отзывов



TF-IDF + Stacking

Векторизация текста и ансамблевая модель для предварительной обработки



BERT (RuBERT)

Предварительно обученная модель для создания семантических эмбеддингов



Logistic Regression

Финальный классификатор для определения тональности текста



PCA

Метод понижения размерности для визуализации кластеров



WordCloud, Plotly, Matplotlib

Библиотеки для создания визуализаций и графиков



Streamlit

Фреймворк для создания интерактивного веб-приложения

Качество предсказания по классам на размеченной выборке

Точность модели по классам тональности

correct_match	False	True	Всего	Точность %
Негатив	285	9715	10000	97.15%
Нейтрально	570	7920	8490	93.29%
Позитив	385	9615	10000	96.15%



Визуализация точности модели по классам

Особенности реализации модели

🔗 Модель

Используется **комбинированный подход** — TF-IDF + Stacking + BERT. Финальный классификатор принимает на вход как вероятности от стекинг-модели, так и CLS-эмбеддинги из BERT.

👤 Описание архитектуры

1 TF-IDF-векторизация текста

Преобразование текста в числовые векторы на основе частоты слов

2 Stacking-модель (бустинг + логрессия)

Промежуточное представление для извлечения признаков

3 BERT (RuBERT)

Выделение семантических признаков (CLS-эмбеддинг)

4 Объединение признаков

Финальный классификатор (MLP / логрессия)

🔗 [Архитектура, обучение, предсказание модели: ссылка](#)

Архитектура комбинированной модели





Интерфейс: Ввод и анализ одного отзыва

Функциональность

Приложение позволяет пользователю ввести отзыв и получить:

 Предсказанную тональность

 Уровень уверенности

 График уверенности

 [Запуск веб-интерфейса в streamlit: ссылка](#)

Пример интерфейса анализа отзыва

Отличный сервис! Быстрая доставка и качественный товар. Рекомендую!

Анализировать

 Тональность: Позитивная

 Уверенность модели:

85%


 Распределение вероятностей:


 Визуализация распределения тональностей

Фильтрация данных

Фильтрация отзывов

Пользователь может фильтровать отзывы:

 По дате

 По типу организации

Результаты анализа

 Метки тональности

 Уверенность

 Вероятности

Интерфейс фильтрации данных

Период

01.01.2023



—

31.12.2023



Тип организации

Все организации






Количество отзывов для анализа

100

ОТЗЫВОВ

Применить фильтры

Предпросмотр результатов

-  Отличный сервис, очень доволен... 92%
-  Плохое качество, не рекомендую... 87%
-  Обычное обслуживание, ничего особенного... 75%

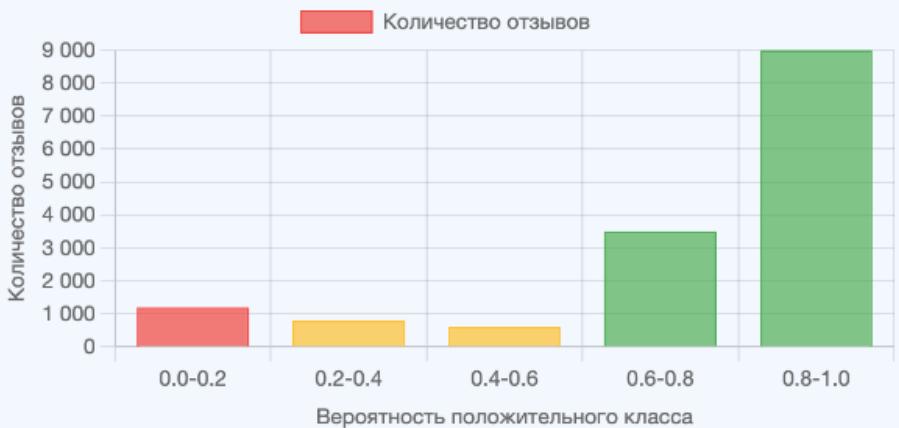
Распределение по тональности и вероятности

ii Гистограмма вероятностей
положительного класса

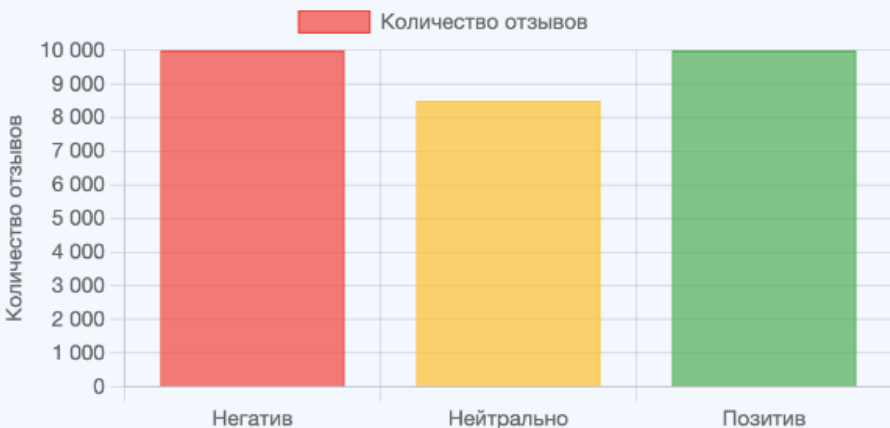
ii Столбчатая диаграмма
количества по меткам

⌵ Отображение отзывов с
низкой уверенностью (менее
65%)

ii. Гистограмма вероятностей



iii. Распределение по тональностям



⌵ Отзывы с низкой уверенностью



⚡ Анализ уверенности модели



💡 Если пики ближе к 0.5, модель часто сомневается; если ближе к 0 или 1 — она уверена.

Кластеризация с помощью PCA

Отзывы отображаются в 2D-пространстве



Эмбединги сгенерированы через BERT

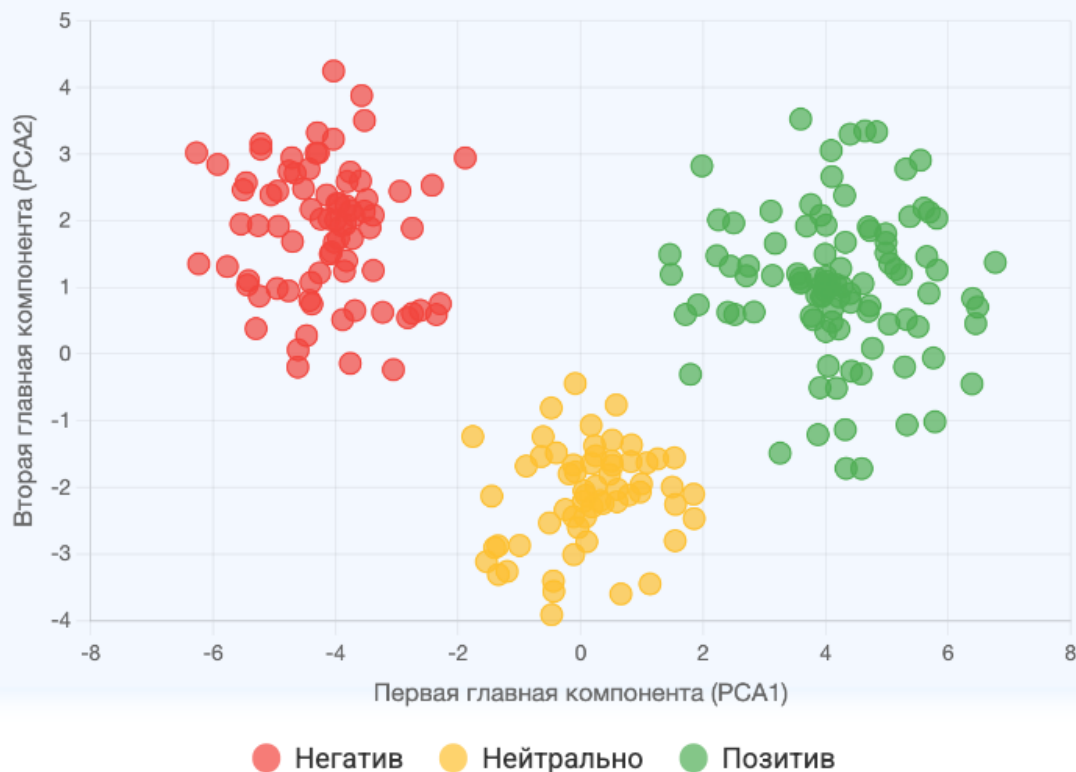


Уменьшение размерности через PCA



Цвета — по тональности

Визуализация кластеров отзывов



Облака слов по тональности



Для каждой тональности — облако слов



Подпись: топ-5 часто встречающихся слов



Автоматически фильтруются короткие и нерелевантные слова



Позитивные отзывы

хорошо быстро
удовно довольный
качественно рекомендую
спасибо

Топ-5 слов:

отлично хорошо
рекомендую качественно
быстро



Негативные отзывы

ужасно дорого
плохо разочарован
неудобно медленно
рекомендую
проблема

Топ-5 слов:

плохо ужасно
не рекомендую медленно
дорого



Нейтральные отзывы

средне типично
нормально приемлемо
ожидаемо обычно
стандартно
нейтрально

Топ-5 слов:

нормально средне обычно
стандартно типично

Технические сложности и решения

Преодоленные технические вызовы в процессе разработки

Проблема с BERT

BERT не предназначен для прямой интеграции с классическими ML-моделями

Решение

Требовалась обёртка, обеспечивающая батчинг, паддинг, работу с GPU/CPU

Конкатенация больших тензоров

Объединение BERT-эмбеддингов и вероятностей стек-модели приводило к высоким требованиям к памяти

Решение

Решено путём:

- ✓ Использования батчей при генерации эмбеддингов
- ✓ Ограничения на `max_length=128`
- ✓ Работы с `float32` вместо `float64` для уменьшения размера

Модель неуверенна

Некоторые прогнозы модели имели низкую уверенность

Решение

Указан уровень доверия и отдельный вывод слабых предсказаний

Результат

★ Достижения проекта

✓ **Интерактивное приложение** ГОТОВО К ИСПОЛЬЗОВАНИЮ

📝 Поддерживает **массовую обработку** и визуальный анализ

⚙️ **Гибкая архитектура:** можно адаптировать под любую предметную область (отзывы, сообщения и т.д.)

Предпросмотр приложения

📊 Анализ текстовых отзывов

📈 Статистика анализа

📊 Обработано отзывов: 28,490

👍 Позитивных: 35.1%

👎 Негативных: 35.1%

— Нейтральных: 29.8%

🔗 Визуализация кластеров



2D-визуализация кластеров отзывов

Дальнейшие шаги

Планы развития проекта



Подключение к базе данных

Интеграция с базой данных отзывов в **реальном времени** для актуального анализа



Обучение по разметке экспертов

Добавление возможности **дообучения** модели на основе новой разметки от экспертов



Улучшение визуализации

Внедрение продвинутых методов **визуализации** (UMAP, t-SNE) для более точного представления данных



Поддержка английского языка

Расширение функционала для **многоязычного** анализа текстов

Дорожная карта развития



Интеграция с базой данных
Ближайшие 3 месяца



Система дообучения модели
3-6 месяцев



Новые методы визуализации
6-9 месяцев



Мультиязычная поддержка
9-12 месяцев