Классификация эмоций в текстовых расшифровках голосовых сообщений

КОМАНДА

Бадретдинова Рушана	Тим Лид
Евдокимов Денис	Участник команды
Белова Виктория	Участник команды
Комаревцева Анна	Участник команды
Артюшев Рафаэль	Участник команды

ЗАДАЧА

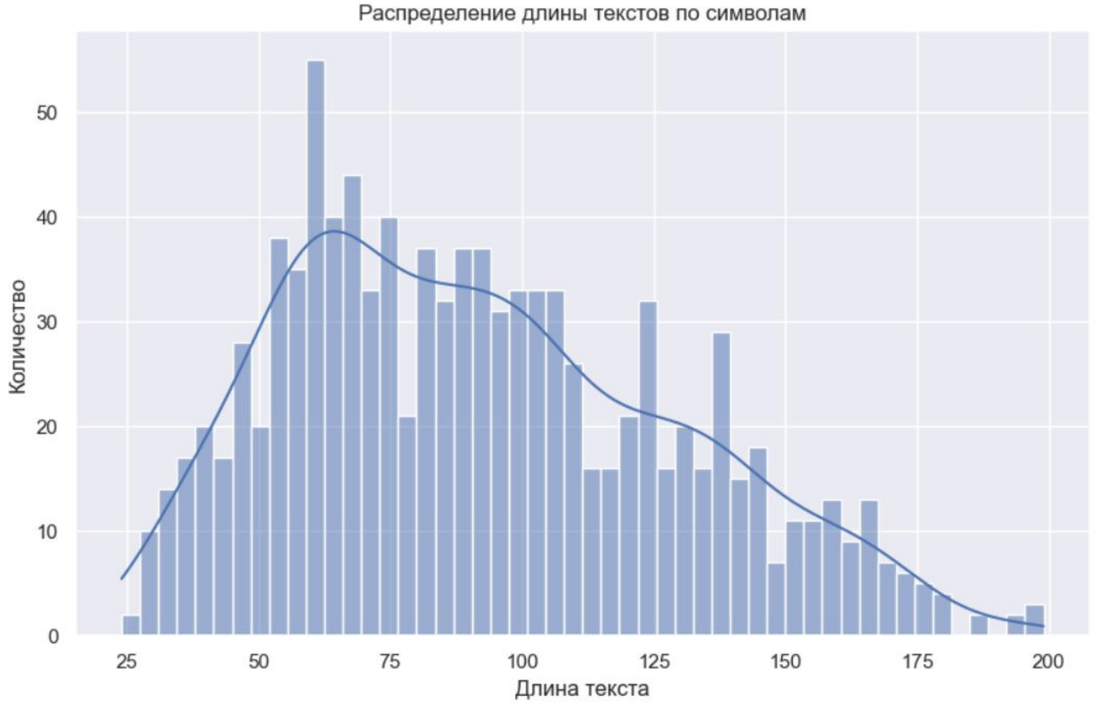
Определить эмоциональное состояние собеседника, проводя анализ текстовой расшифровки его голосового сообщения.

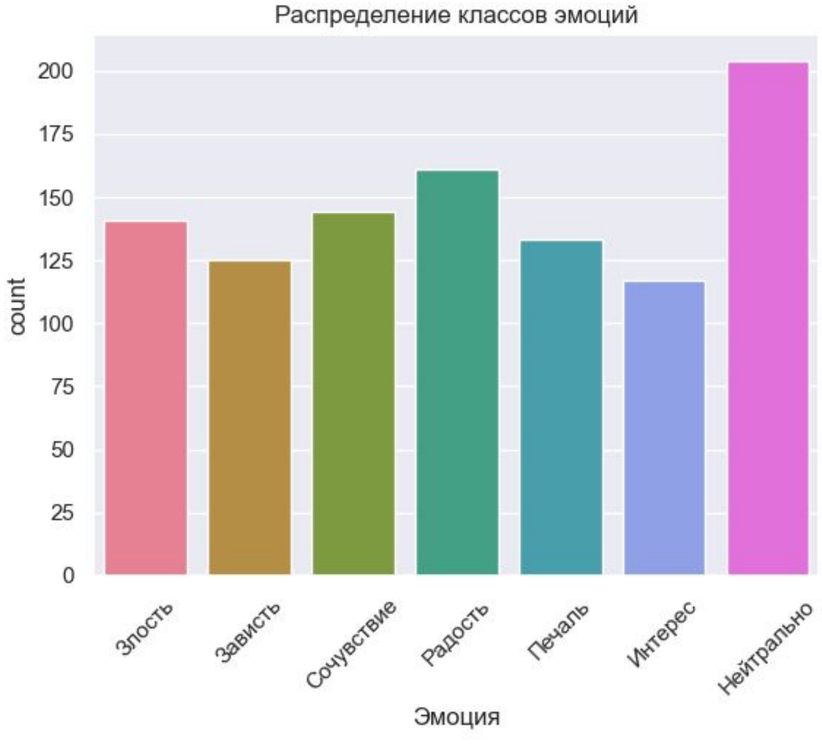
- Предварительная предобработка данных: нормализация, токенизация, удаления стоп слов и т.д.
- Построение модели, которая обладает устойчивостью к разнообразию данных и может ясно их интерпретировать.

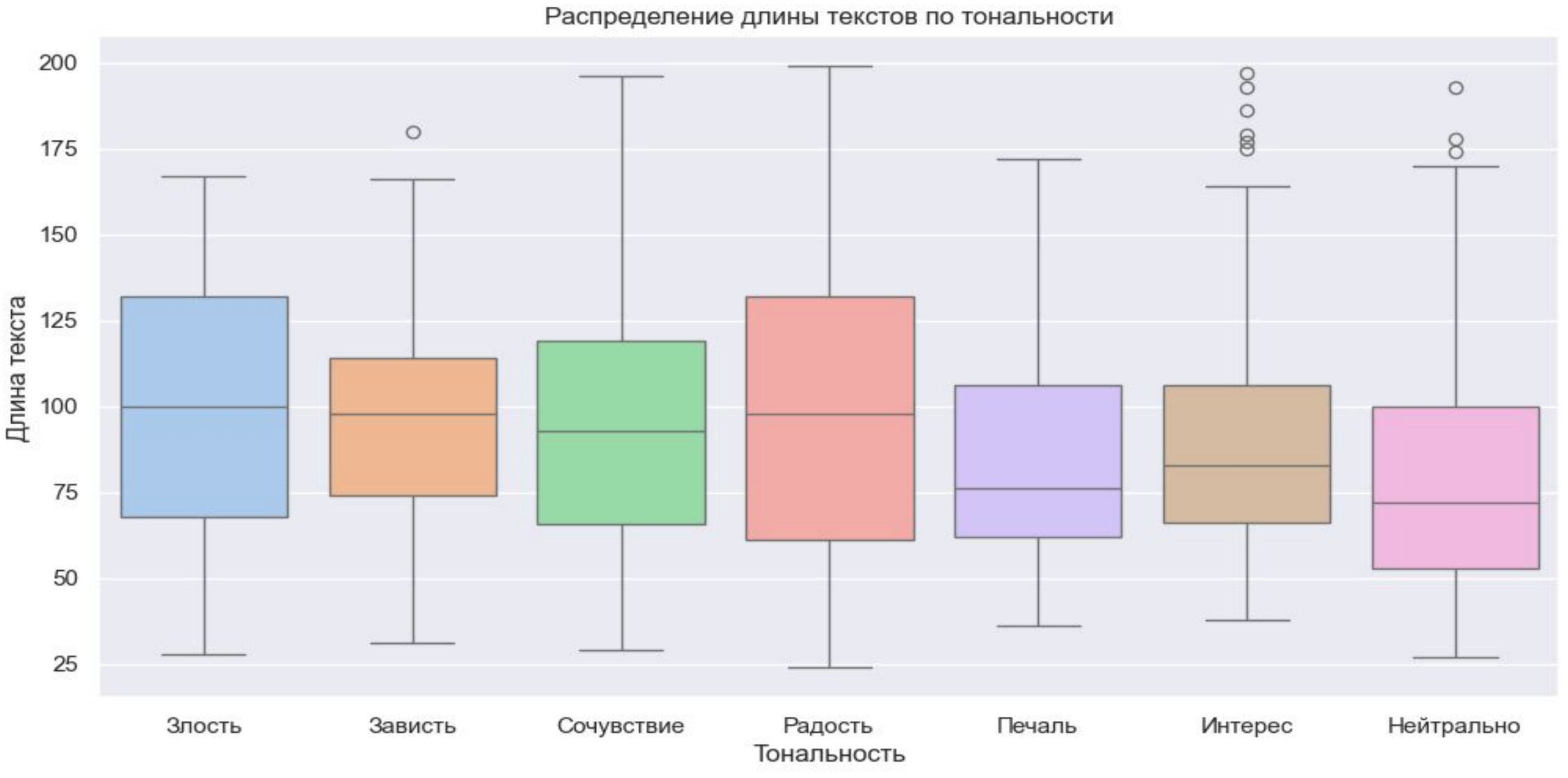
ПРЕДОБРАБОТКА ДАННЫХ

- Удаление дубликатов, пропусков
- Удаление класса “Недовольство”, так как этот класс плохо представлен.
- Выявление самых длинных текстов и их обрезка, так как длинные тексты скажутся на качестве модели.
- Нормализация:
  - Приведение к нижнему регистру, удаление пунктуации, цифр
  - удаление стоп-слов с помощью библиотеки nltk
  - стемминг с помощью SnowballStemmer
  - токенизация - CountVectorizer, Word2Vec

EXPLORATION DATA ANALYSIS

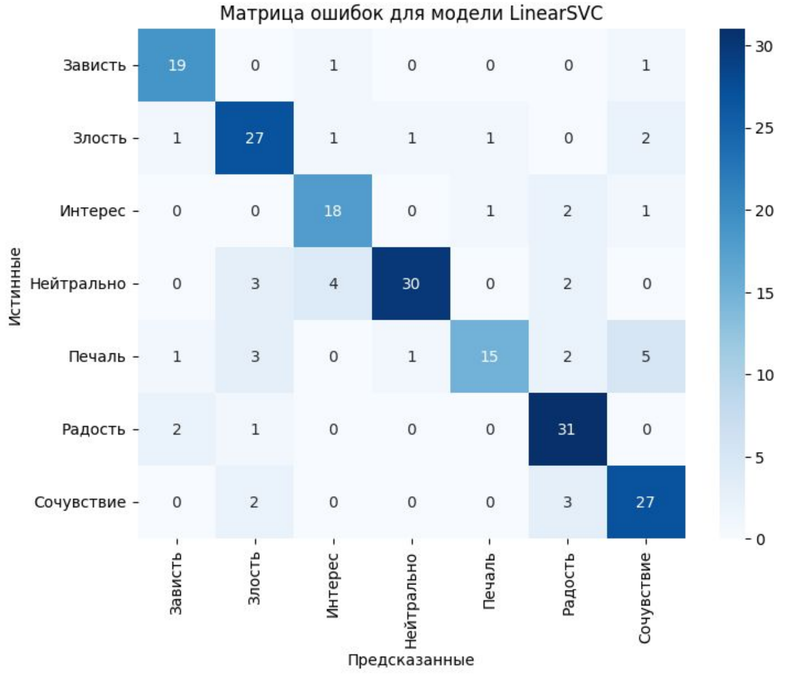





Распределение длины текстов по тональности
Длина текста
200
175
150
125
100
75
50
25
Злость
Зависть
Сочувствие
Радость
Печаль
Интерес
Нейтрально
Тональность

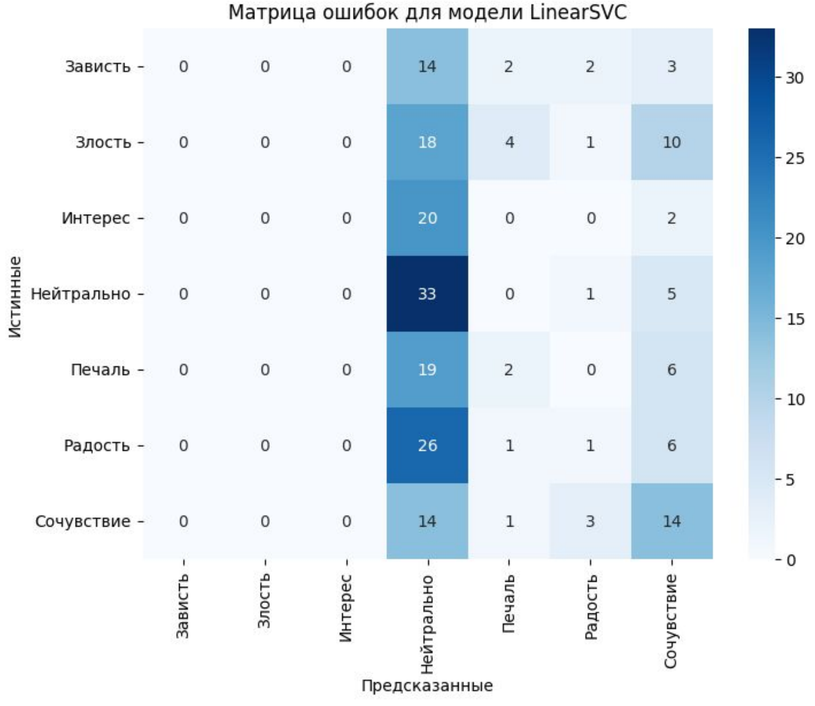
Метрики классических моделей машинного обучения. Предобработка CountVectorizer.

Модель	Среднний F1-score	Средний Precision	Средний Recall
наивный Байес	0.74	0.80	0.75
Логистическая Регрессия	0.73	0.76	0.74
LinearSVC	0.80	0.80	0.81



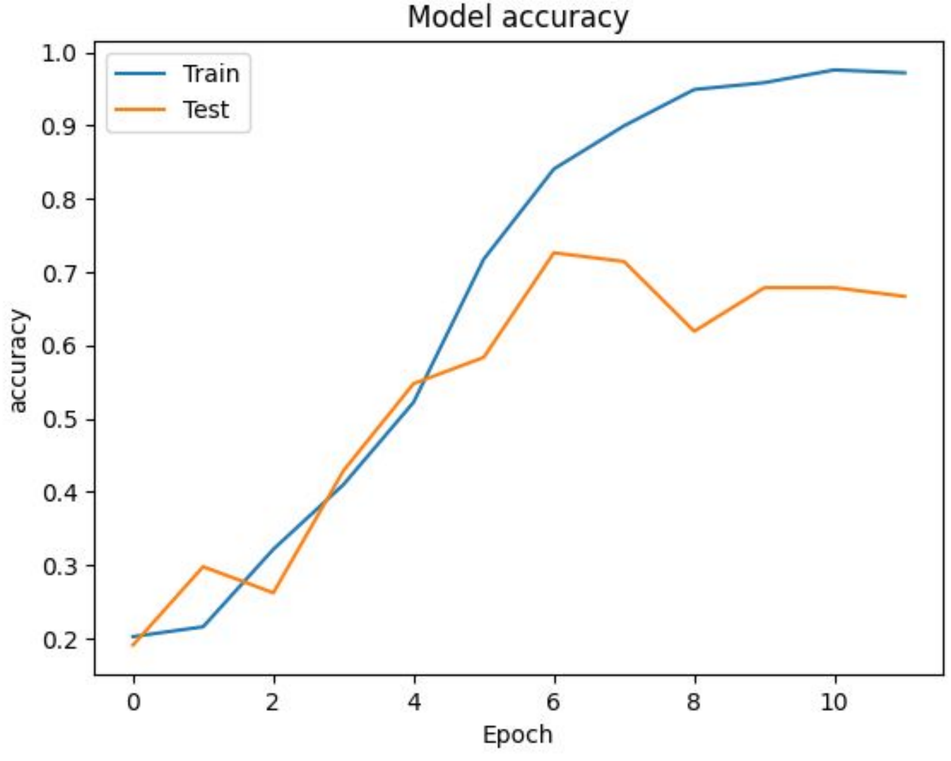
Метрики классических моделей машинного обучения. Предобработка **Word2Vec.**

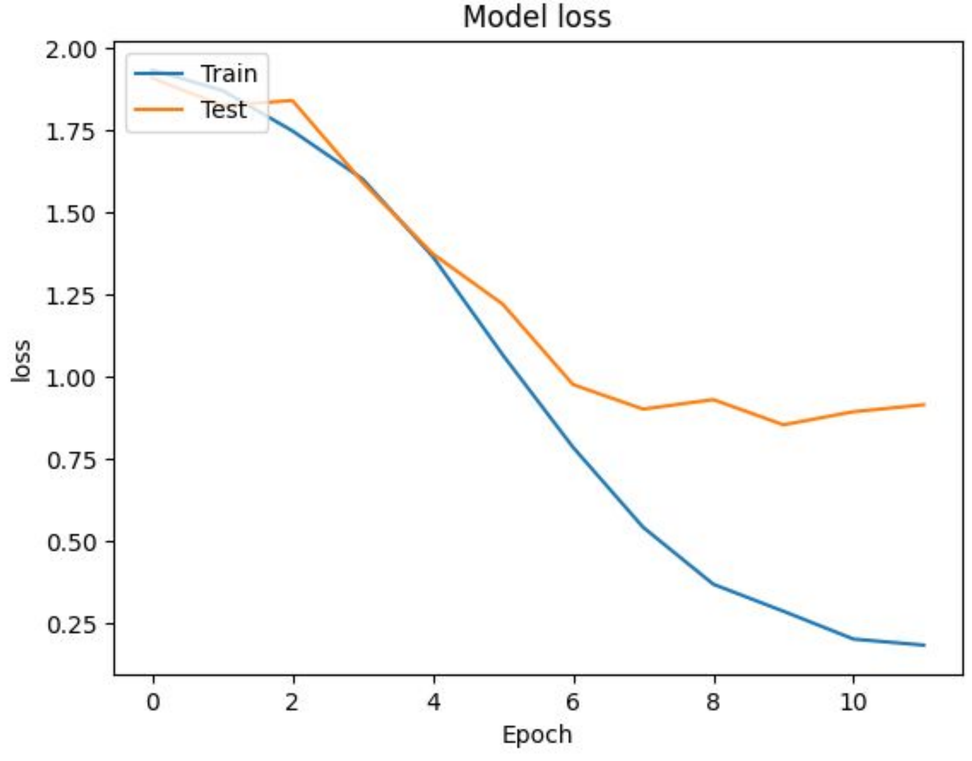
Модель	Средний Precision	Средний Recall	Средний F1-score
Наивный Байес	0.23	0.27	0.22
Логистическая Регрессия	0.14	0.24	0.14
LinearSVC	0.14	0.24	0.14



Bi-LSTM.Tokenizer из библиотеки Keras

Параметры:
Количество эпох = 12
Размер батча = 32
Validation_split = 0.1
Количество слоев = 3





МФТИ

Спасибо за внимание!