

Исследование тональности отзывов покупателей онлайн-магазина одежды

Выполнил: Гребенюк А.В.

Научный руководитель: Семендяев Р.Ю.

Санкт-Петербург
2023

Цель работы: определение оптимального инструмента для анализа тональности отзывов покупателей онлайн-магазина

- Обзор основных подходов к исследованию тональности текста
- Обзор и тестирование существующих предобученных моделей
- Определение оптимальной модели для обучения на выбранном датасете

► **Анализ тональности текста** (*Sentiment analysis*) — класс методов контент-анализа в компьютерной лингвистике и NLP, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов по отношению к объектам, речь о которых идёт в тексте.

- **Классификация по бинарной шкале**
- **Классификация по многополосной шкале**

Описание дата-сета

Признаки:

1. Возраст автора отзыва
2. Заголовок отзыва
3. Текст отзыва
4. Рейтинг товара по шкале от 1 до 5
5. Рекомендация «да/нет»
6. Количество отметок «нравиться» у отзыва

Объем датасета: 23486 записей

```
RangeIndex: 23486 entries, 0 to 23485
```

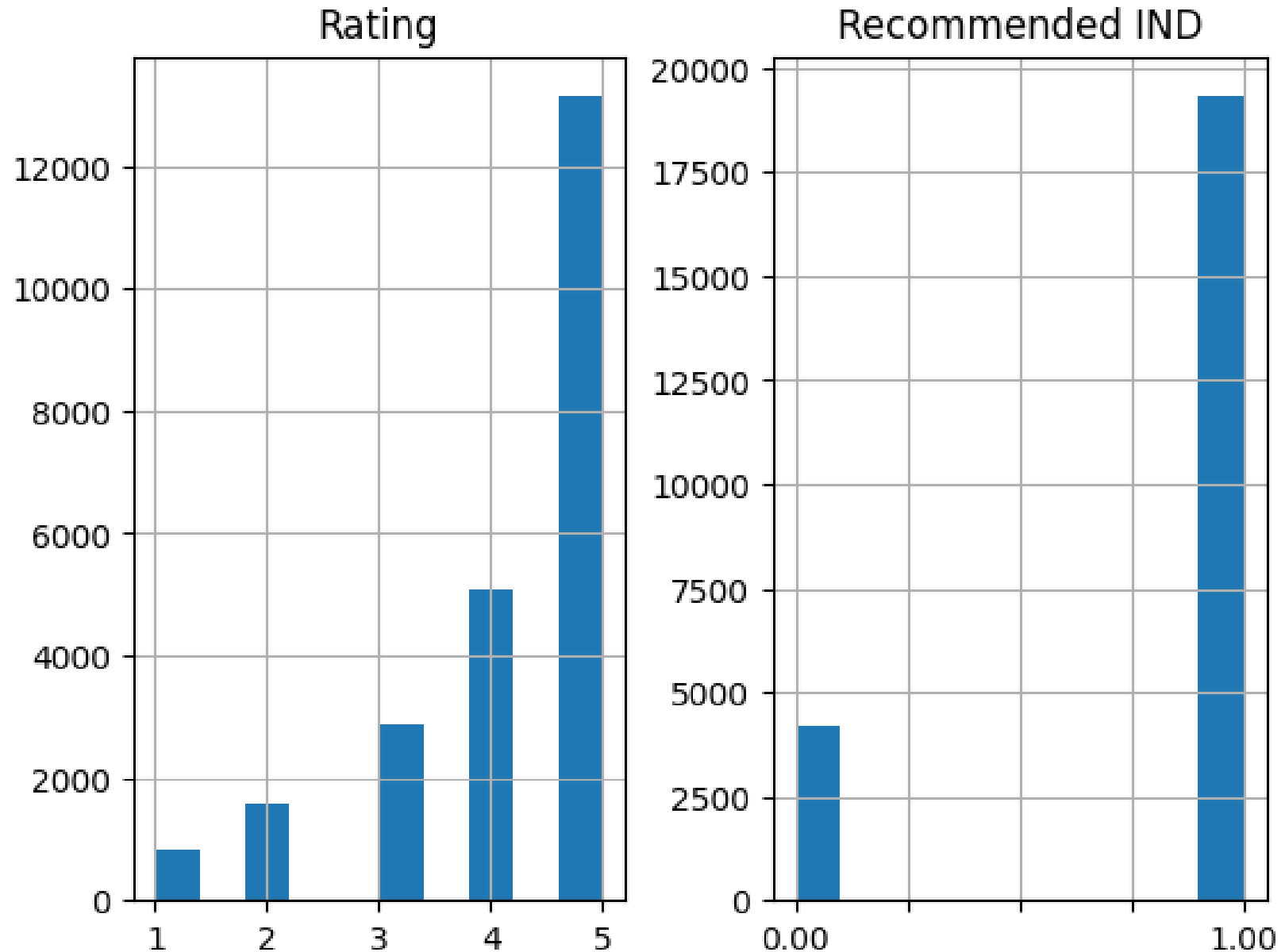
```
Data columns (total 11 columns):
```

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	23486 non-null	int64
1	Clothing ID	23486 non-null	int64
2	Age	23486 non-null	int64
3	Title	19676 non-null	object
4	Review Text	22641 non-null	object
5	Rating	23486 non-null	int64
6	Recommended IND	23486 non-null	int64
7	Positive Feedback Count	23486 non-null	int64
8	Division Name	23472 non-null	object
9	Department Name	23472 non-null	object
10	Class Name	23472 non-null	object

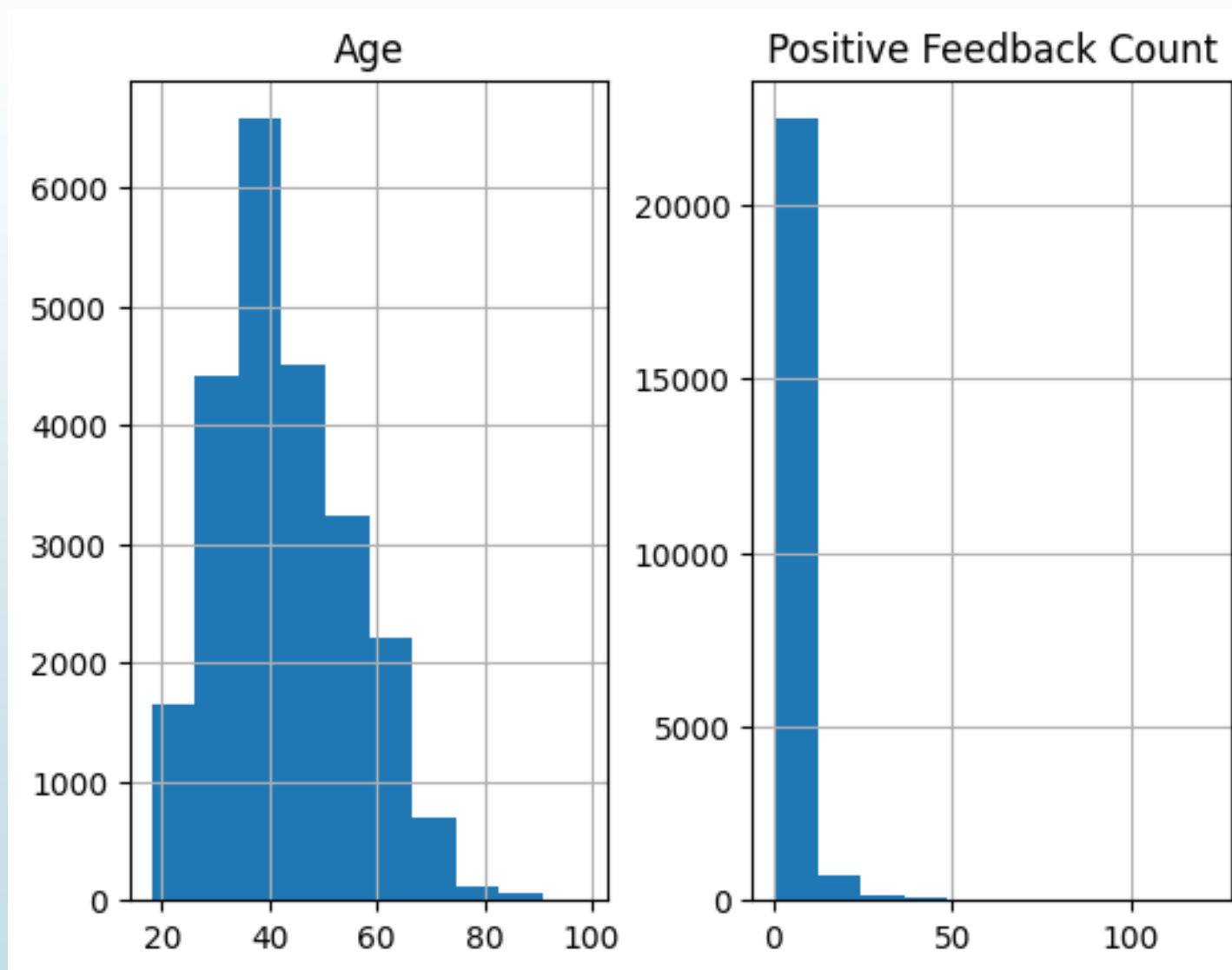
```
dtypes: int64(6), object(5)
```

Распределение оценок и рекомендаций:

5

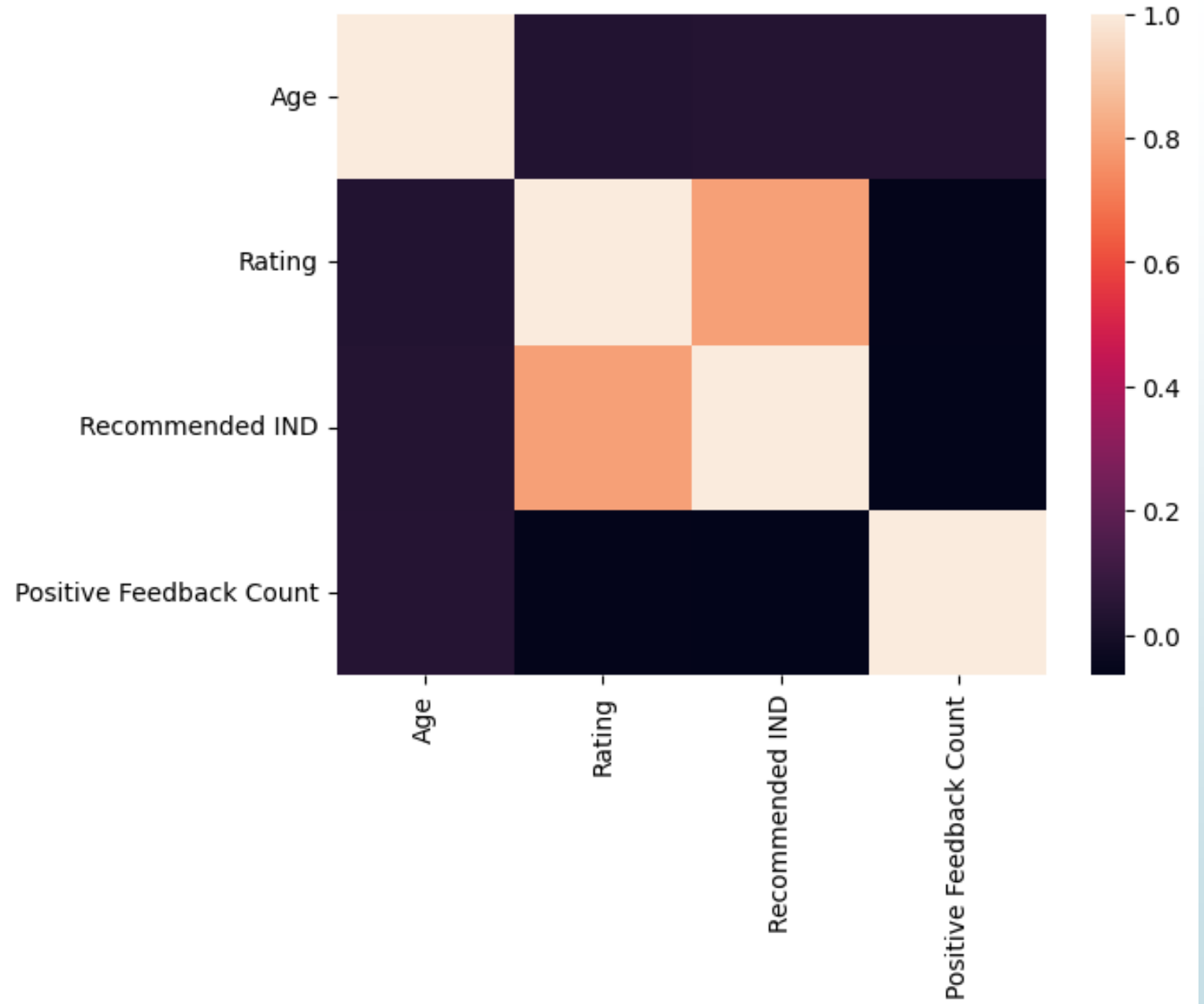


Распределение возрастов и отметок «нравиться» у отзывов



7

Корреляция признаков:



Использованные модели:

- Предобученные модели (TextBlob, Bert, Hugging Face)
- Модели классического ML (LR, KNN, DTC, SVC, XGBoost)
- Нейросеть LSTM:

```
model = Sequential()
model.add(Embedding(maxWordsCount, 128, input_length = max_text_len))
model.add(LSTM(128, return_sequences=True))
model.add(LSTM(64))
model.add(Dense(2, activation='softmax'))
model.summary()

model.compile(loss='categorical_crossentropy', metrics=['accuracy'], optimizer=Adam(0.0001))

history = model.fit(X, Y, batch_size=32, epochs=50)
```

Предобработка данных

- Графематический анализ (токенизация, удаление знаков препинания и стоп-слов)
- Морфологический анализ (лемматизация, присвоение тегов POS)
- Цифровое кодирование слов:
 - Простое кодирование «слово-частотность»
 - Мешок слов
 - Word2Vec – кодирование

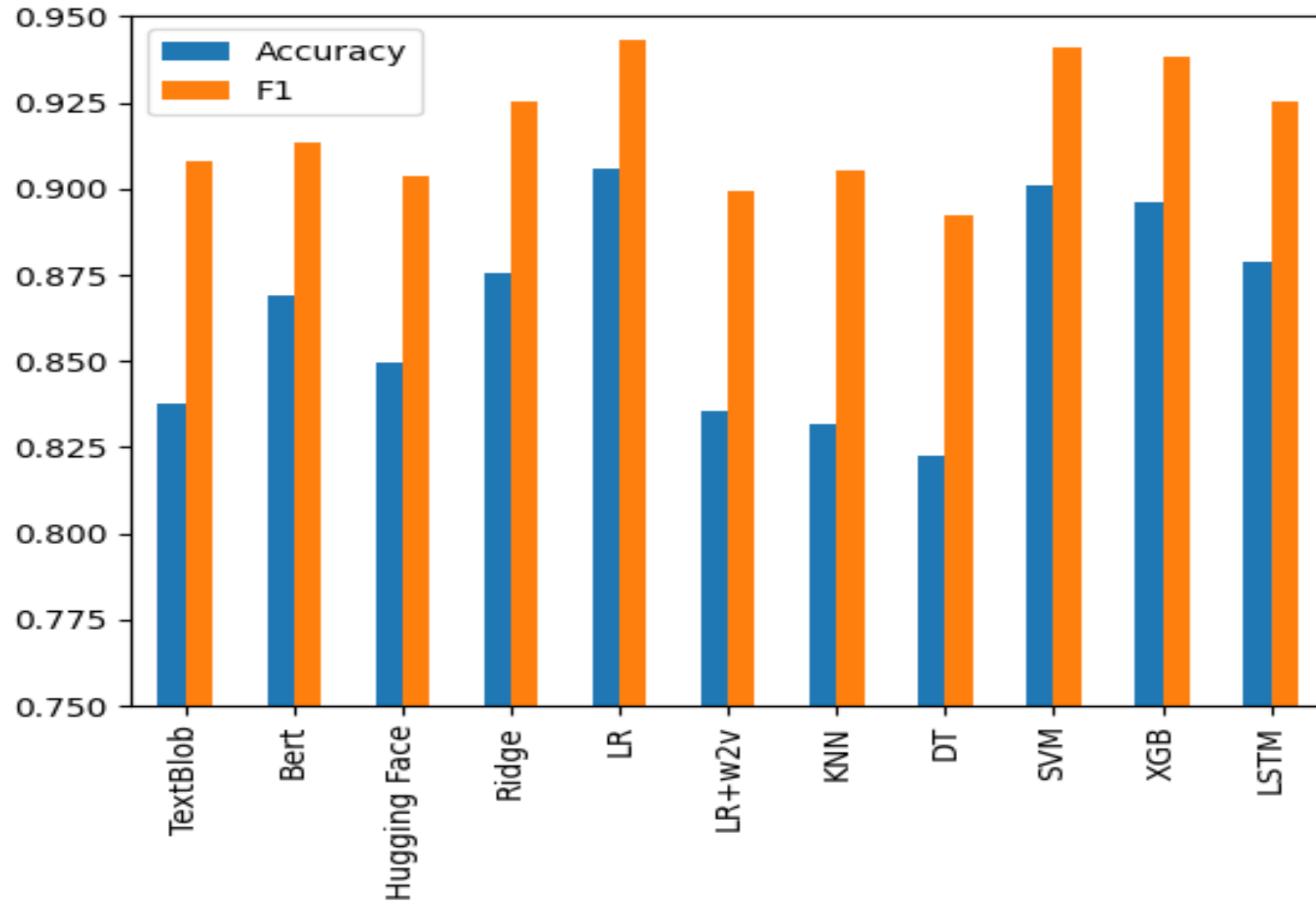
['Я люблю обработку языка']

['Я', 'люблю', 'обработку', 'языка']

['Я', 'любить', 'обработка', 'язык']

[53, 134, 20, 56]

Задача бинарной классификации: рекомендация «да/нет»



Матрицы ошибок:

Модель TextBlob

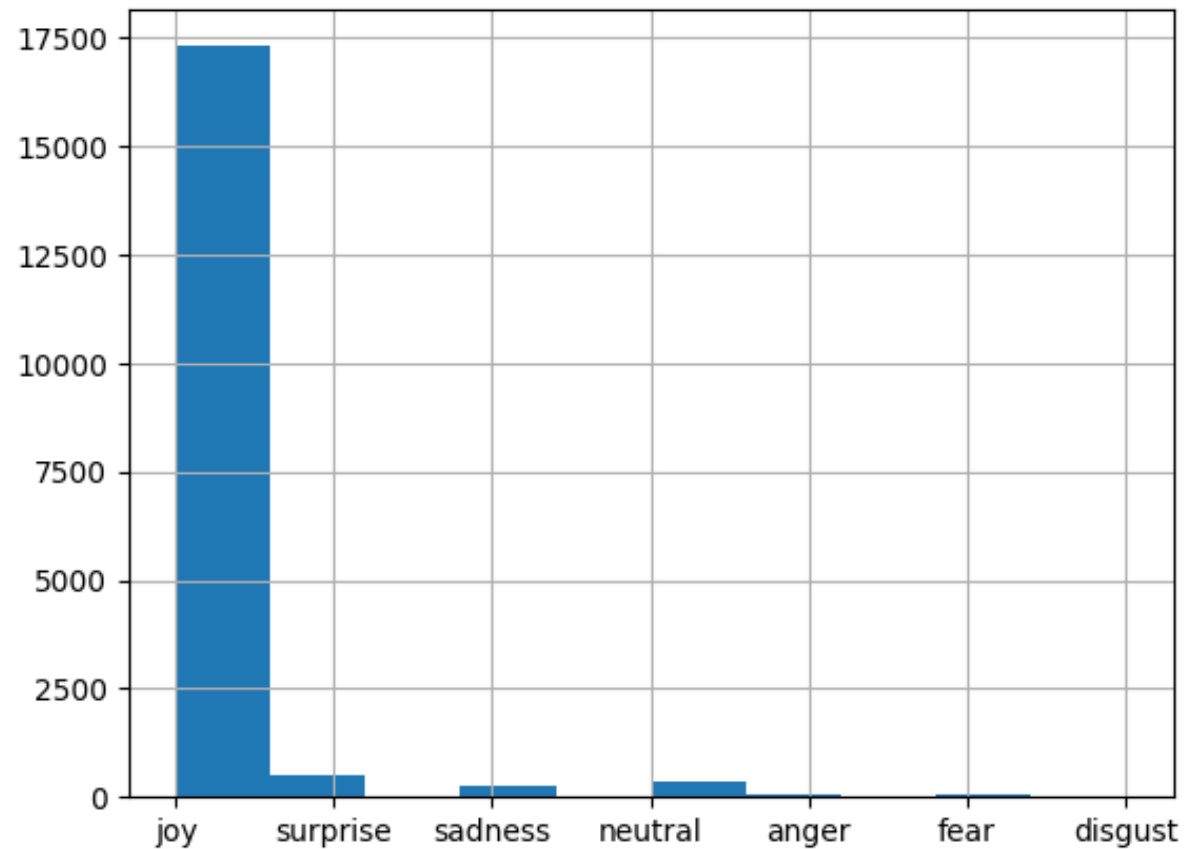
	Act Negative	Act Positive
Pred Negative	804	377
Pred Positive	3297	18163

Модель LSTM

	Act Negative	Act Positive
Pred Negative	564	301
Pred Positive	248	3416

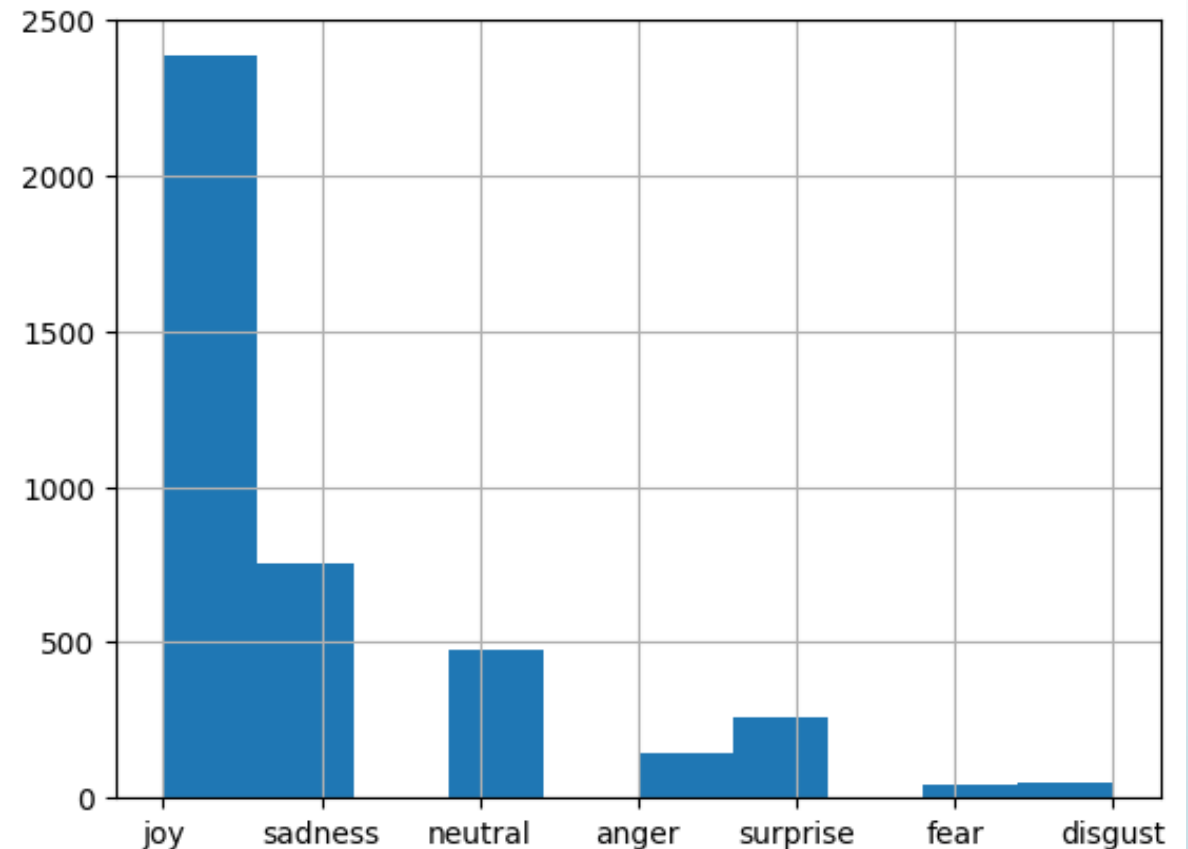
Доминирующие эмоции положительных и отрицательных отзывах

Положительные отзывы



Токсичность: 39%

Отрицательные отзывы



Токсичность: 43%

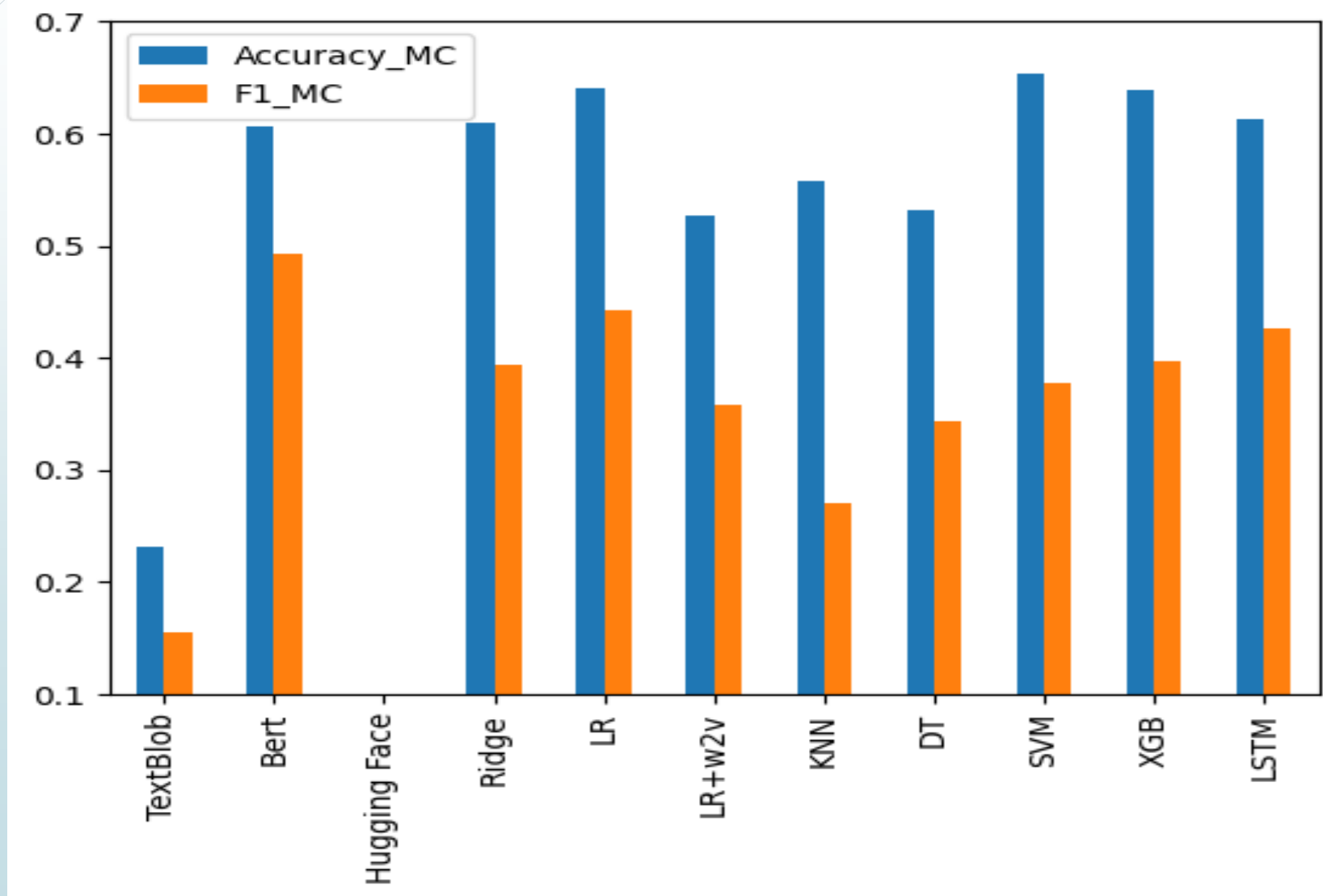
Наиболее частотные существительные и прилагательные

13

	Существительные		Прилагательные в положительных отзывах	Прилагательные в отрицательных отзывах
1	dress	14	good	thin
2	size	16	true	little
3	color	17	super	disappointed
4	top	18	short	soft
5	fabric	19	big	wide
6	shirt	20	flattering	tight
7	sweater	21	white	high
8	bit	22	gorgeous	boxy
9	skirt	23	casual	bad
10	material	24	petite	white
11	length	25	long	cheap
12	quality	26	unique	first
13	pants	27	easy	huge
14	jeans	28	loose	poor
15	waist	29	tight	unflattering

Задача множественной классификации: оценка от 1 до 5

14



Выводы:

- ▶ При отсутствии семантического кодирования (word2vec) нейронные сети LSTM не дают существенного преимущества по сравнению с предобученными моделями и моделями классического машинного обучения
- ▶ Лучшее решение задачи бинарной классификации по метрике F1 продемонстрировала логистическая регрессия
- ▶ Лучшее решение задачи многоклассовой классификации по метрике F1 продемонстрировала предобученная модель “Bert”