Test tapşırığını rus dilində yerinə yetirdim (üzr istəyirəm, amma hələlik bu cür düşünmək mənim üçün daha asandır), amma hesabatı azərbaycan dilində təkrarladım.

ОТЧЁТ ПО ТЕСТОВОМУ ЗАДАНИЮ.

1. Изучение анализа тональности

Перед тем как начать, я изучил, что такое анализ тональности текста и какие методы существуют.

Я понял, что анализ тональности – это определение эмоциональной окраски текста. Иногда это просто положительное, отрицательное или нейтральное настроение, но в нашем проекте задача была глубже – определить конкретную эмоцию, такую как радость, грусть, гнев, страх, любовь или удивление.

Я рассмотрел несколько подходов:

1) Лексиконный метод. Он работает на основе словарей, где каждому слову присвоена эмоция. Пример: слово «счастливый» – это радость. Этот метод простой, но не учитывает контекст, поэтому часто даёт ошибки.
2) Метода машинного обучения. Они строят модели на размеченных данных, учат связь между словами и эмоциями. Этот подход более точный и масштабируемый.
3) Глубокое обучение (нейросети, LSTM, BERT). Он позволяет учитывать контекст и сложные зависимости в тексте. Работает очень хорошо, но требует много ресурсов и времени для обучения.

Я решил использовать машинное обучение с TF-IDF и логистической регрессией, потому что это даёт хороший баланс точности и простоты, и результаты легко объяснить.

2. Загрузка набора данных и предварительный анализ (EDA)

Для начала работы с набором данных я загрузил 3 файла: тренировочный (training.csv), валидационный (validation.csv), тестовый (test.csv). В тренировочном наборе содержится 16000 сообщений, валидационный и тестовые наборы содержат по 2000 сообщений каждый. Каждое сообщение имеет 2 поля: text – текст, и label – числовая метка эмоции от 0 до 5.

Я проверил первые записи и убедился, что данные корректные, все строки заполнены, пропусков нет. Примеры текстов показывают, что сообщения разные по длине и эмоциональному содержанию, что подходит для обучения модели классификации эмоций.

Далее я изучил распределение эмоций в тренировочном наборе. Оно оказалось несбалансированным:

1) Наиболее часто встречаются эмоции страх и гнев;
2) Средняя частота у эмоций любовь и печаль;
3) Редко встречаются радость и удивление.

Такое распределение важно учитывать при построении модели, чтобы она не была смещена в сторону более частых эмоций.

Я также проанализировал длину текстов. Средняя длина сообщений составляет 97 символов, медиана – 86 символов. Большинство сообщений короткие или средней длины, длинные тексты встречаются реже.

Это подтверждает, что использование TF-IDF векторизации подходит для представления текстов, так как она хорошо работает с небольшими и средними по длине текстовыми данными.

Визуализация распределения эмоций и длины текстов показала, что:

1) Эмоции fear и anger занимают почти половину всего тренировочного набора;
2) Эмоции joy и surprise встречаются реже;
3) Большинство сообщений короткие, с пиком длины около 50–80 символов.

Выводы по EDA:

1) Данные корректные, пропусков нет;
2) Распределение эмоций несбалансированное, что стоит учитывать при обучении;
3) Длины сообщений подходит для TF-IDF векторизации и обучения модели;
4) Набор данных готов для построения модели классификации эмоций.

```
PS D:\Project\task_emotion> py src/eda.py
TRAIN SHAPE: (16000, 2)
VALIDATION SHAPE: (2000, 2)
TEST SHAPE: (2000, 2)

TRAIN COLUMNS:
Index(['text', 'label'], dtype='object')

Первые 5 строк:
                                                text  label
0                            i didnt feel humiliated      0
1  i can go from feeling so hopeless to so damned...      0
2   im grabbing a minute to post i feel greedy wrong      3
3  i am ever feeling nostalgic about the fireplac...      2
4                                i am feeling grouchy      3

Пропущенные значения:
text     0
label    0
dtype: int64

Распределение эмоций:
emotion
fear        5362
anger       4666
love        2159
sadness     1937
joy         1304
surprise     572
Name: count, dtype: int64

Статистика длины текста:
count    16000.000000
mean        96.845812
std         55.904953
min          7.000000
25%         53.000000
50%         86.000000
75%        129.000000
max        300.000000
Name: text_length, dtype: float64
```

3. Обработка текста

Прежде чем обучать модель, я провёл предобработку текста. Это важно, потому что машинное обучение лучше работает с чистыми и стандартизированными данными.

Я сделал следующие шаги:

1) Преобразовал все буквы текста в строчные. Это помогает модели не различать, например, «Счастливый» от «счастливый»;
2) Удалил все цифры, так как они обычно не несут эмоции и могут только шумить модель;
3) Удалил пунктуацию и лишние пробелы.

После этих шагов тексты стали единообразными и подготовленными для TF-IDF векторизации.

```python
import re

def clean_text(text: str) -> str:
    text = text.lower()
    text = re.sub(r'\d+', '', text)       # удаление чисел
    text = re.sub(r'[^\w\s]', '', text)   # удаление пунктуации
    text = text.strip()
    return text
```

4. Распределение эмоций в тренировочном наборе

Emotion	**Количество**
fear	5362
anger	4666
love	2159
sadness	1937
joy	1304
surprise	572

```
Распределение эмоций:
emotion
fear        5362
anger       4666
love        2159
sadness     1937
joy         1304
surprise     572
```

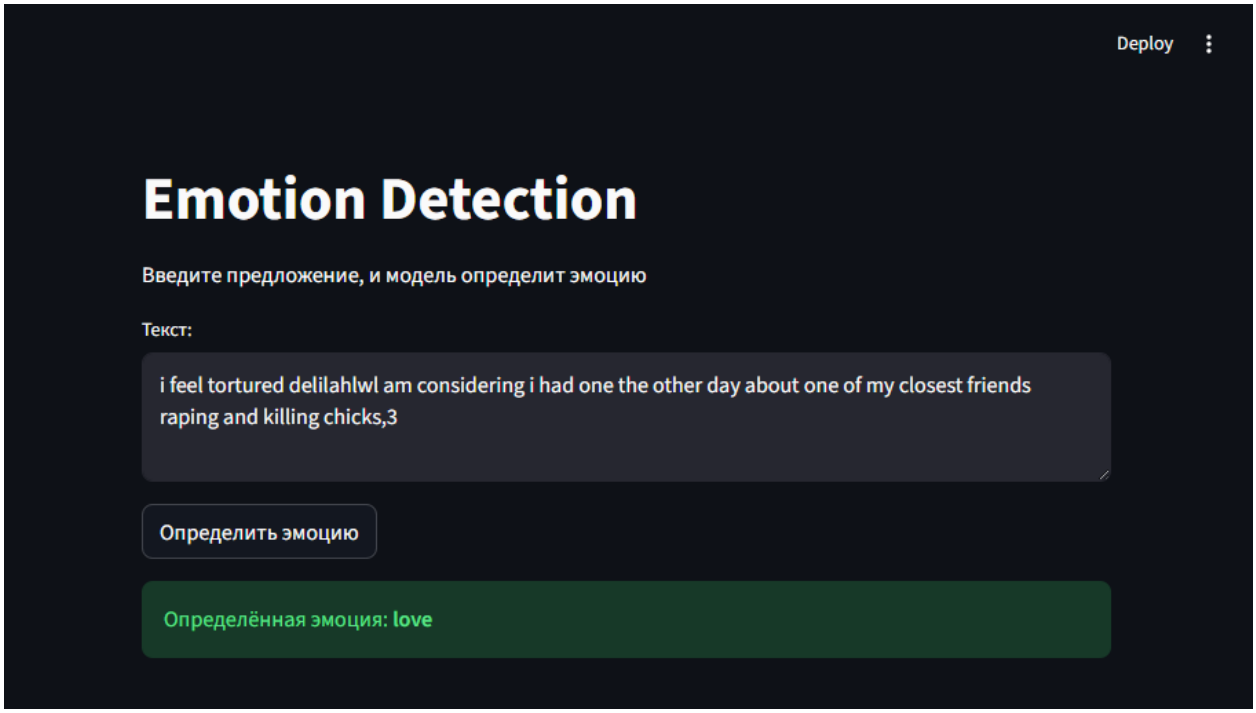
Вывод:

1) Данные несбалансированы, fear и anger встречаются чаще, а joy и surprise реже;
2) Это нужно учитывать при обучении, так как модель может быть склонна предсказывать более частые эмоции.

5. **Длина текстов**

```
Статистика длины текста:
count    16000.000000
mean        96.845812
std         55.904953
min          7.000000
25%         53.000000
50%         86.000000
75%        129.000000
max        300.000000
```

6. **Создайте простой интерфейс с помощью Gradio или Streamlit и проверьте производительность модели, создав собственные примеры для каждой эмоции.**



TEST TAPŞIRIĞI HESABATI.

1. Hiss təhlilinin öyrənilməsi

Başlamazdan əvvəl duyğu təhlilini və onun metodlarını öyrəndim.

Duyğu təhlilinin mətnin emosional tonunun müəyyən edilməsi olduğunu başa düşdüm. Bəzən bu, sadəcə müsbət, mənfi və ya neytral bir duyğu olur, lakin layihəmizdə məqsəd daha dərin idi – sevinc, kədər, qəzəb, qorxu, sevgi və ya təəccüb kimi müəyyən bir duyğunu müəyyən etmək.

Bir neçə yanaşmanı nəzərdən keçirdim:

1) Leksikona əsaslanan metodlar. Bu, hər sözə bir duyğu təyin edildiyi lüğətlərə əsaslanır. Məsələn, "xoşbəxt" sözü sevincdir. Bu metod sadədir, lakin konteksti nəzərə almır, buna görə də tez-tez səhvlərə yol verir.
2) Maşın öyrənmə metodları. Bu metodlar etiketli məlumatlar üzərində modellər qurur, sözlər və duyğular arasındakı əlaqəni öyrənir. Bu yanaşma daha dəqiq və miqyaslıdır.
3) Dərin öyrənmə (neyron şəbəkələri, LSTM, BERT). Bu, mətndə kontekst və mürəkkəb asılılıqlara imkan verir. Çox yaxşı işləyir, lakin təlim üçün çoxlu resurs və vaxt tələb olunur. Dəqiqlik və sadəlik arasında yaxşı bir tarazlıq yaratdığı və nəticələrin izahı asan olduğu üçün maşın öyrənməsini TF-IDF və logistik reqressiya ilə istifadə etməyə qərar verdim.

2. Məlumat dəstinin yüklənməsi və ilkin təhlil (EDA)

Məlumat dəsti ilə işləməyə başlamaq üçün üç fayl yüklədim: təlim (training.csv), validation (validation.csv) və test (test.csv). Təlim dəstində 16.000 mesaj, validation və test dəstlərində isə 2.000 mesaj var. Hər mesajın iki sahəsi var: mətn və etiket, 0-dan 5-ə qədər ədədi emosiya etiketi.

İlkin qeydləri yoxladım və məlumatların düzgün olduğunu, bütün sətirlərin doldurulduğunu və boşluqların olmadığını təsdiqlədim. Nümunə mətnlər göstərir ki, mesajlar uzunluq və emosional məzmun baxımından dəyişir ki, bu da emosiya təsnifatı modelini öyrətmək üçün uyğundur.

Sonra təlim dəstindəki emosiyaların paylanmasını araşdırdım. Məlum oldu ki, balanssızdır:

1) Qorxu və qəzəb ən çox yayılmış emosiyalardır;
2) Sevgi və kədər orta dərəcədə tez-tez olur;
3) Sevinc və təəccüb nadirdir.

Model qurarkən daha çox yayılmış emosiyalara meylli olmadığından əmin olmaq üçün bu paylanmanı nəzərə almaq vacibdir. Mətn uzunluqlarını da təhlil etdim. Orta mesaj uzunluğu 97 simvol, median isə 86 simvoldur. Əksər mesajlar qısa və ya orta uzunluqdadır, daha uzun mətnlər isə daha az yaygındır.

Bu, TF-IDF vektorlaşdırmasının mətn təsviri üçün uyğun olduğunu təsdiqləyir, çünki kiçik və orta uzunluqlu mətn məlumatları ilə yaxşı işləyir.

Emosiyaların və mətn uzunluqlarının paylanmasını vizuallaşdırmaq göstərdi ki:

1) Qorxu və qəzəb emosiyaları bütün təlim dəstinin demək olar ki, yarısını təşkil edir;
2) Sevinc və təəccüb emosiyaları daha az yaygındır;
3) Əksər mesajlar qısadır, pik uzunluğu təxminən 50-80 simvoldur.

EDA nəticələri:

1) Məlumatlar düzgündür, heç bir simvol itkin düşməyib;
2) Emosiyaların paylanması balanssızdır, bu da təlim zamanı nəzərə alınmalıdır;
3) Mesaj uzunluqları TF-IDF vektorlaşdırması və model təlimi üçün uyğundur;
4) Məlumat dəsti emosiya təsnifatı modelini qurmağa hazırdır.

```
PS D:\Project\task_emotion> py src/eda.py
TRAIN SHAPE: (16000, 2)
VALIDATION SHAPE: (2000, 2)
TEST SHAPE: (2000, 2)

TRAIN COLUMNS:
Index(['text', 'label'], dtype='object')

Первые 5 строк:
                                                text  label
0                            i didnt feel humiliated      0
1  i can go from feeling so hopeless to so damned...      0
2   im grabbing a minute to post i feel greedy wrong      3
3  i am ever feeling nostalgic about the fireplac...      2
4                                i am feeling grouchy      3

Пропущенные значения:
text     0
label    0
dtype: int64

Распределение эмоций:
emotion
fear        5362
anger       4666
love        2159
sadness     1937
joy         1304
surprise     572
Name: count, dtype: int64

Статистика длины текста:
count    16000.000000
mean        96.845812
std         55.904953
min          7.000000
25%         53.000000
50%         86.000000
75%        129.000000
max        300.000000
Name: text_length, dtype: float64
```

3. Mətn emalı

Modeli öyrətməzdən əvvəl mətni əvvəlcədən emal etdim. Bu vacibdir, çünki maşın öyrənməsi təmiz və standartlaşdırılmış məlumatlarla daha yaxşı işləyir.

Aşağıdakı addımları atdım:

1) Bütün mətni kiçik hərflərə çevirdim. Bu, modelin, məsələn, "Xoşbəxt"i "xoşbəxt"dən ayırd etməsinə mane olur;
2) Bütün rəqəmləri sildim, çünki onlar adətən heç bir emosiya ifadə etmir və modelə yalnız səs-küy əlavə edə bilir;
3) Durğu işarələrini və əlavə boşluqları sildim.

Bu addımlardan sonra mətnlər ardıcıl oldu və TF-IDF vektorlaşdırılması üçün hazırlandı.

```python
import re

def clean_text(text: str) -> str:
    text = text.lower()
    text = re.sub(r'\d+', '', text)       # удаление чисел
    text = re.sub(r'[^\w\s]', '', text)   # удаление пунктуации
    text = text.strip()
    return text
```

4. Təlim dəstində emosiyaların paylanması

Emotion	Miqdar
fear	5362
anger	4666
love	2159
sadness	1937
joy	1304
surprise	572

```
Распределение эмоций:
emotion
fear        5362
anger       4666
love        2159
sadness     1937
joy         1304
surprise     572
```

Nəticə:

1) Məlumatlar balanssızdır; qorxu və qəzəb daha çox yayılmış, sevinc və təəccüb isə daha az yayılmışdır;
2) Model daha tez-tez baş verən emosiyaları proqnozlaşdırmağa meylli ola biləcəyi üçün bu, təlim zamanı nəzərə alınmalıdır.

5. **Mətnlərin uzunluğu**

```
Статистика длины текста:
count    16000.000000
mean        96.845812
std         55.904953
min          7.000000
25%         53.000000
50%         86.000000
75%        129.000000
max        300.000000
```

6. **Gradio və ya Streamlit vasitəsilə sadə interfeys yaradın və özünüzdən hər duyğuya aid nümunə yaradaraq modelin performansını birdaha yoxlayın.**