Отчет

Датасет имеет 50000 отзывов на фильмы с присвоенным рейтингом и тональностью (положительный или отрицательный). 25000 – test, 25000 – train.

1. Подготовка данных:
- Загружаются данные из папки "train" и "test" в таблицы pandas, где каждая строка представляет собой отзыв, рейтинг и метку класса (1 - положительный отзыв, 0 - отрицательный отзыв).
- Производится очистка текстов от знаков препинания и стоп-слов (слова, которые обычно не несут смысловой нагрузки). Предварительно удалили из стоп-слов слова, котрые могут нести отрицательный смсыл, тк это модет повлиять на полярность отзыва. Также приводятся все слова к нижнему регистру и производится лемматизация (приведение слов к их базовой форме).

2. Классификация тональности отзывов:
- Отзывы в обучающей выборке и тестовой выборке векторизуются с помощью TF-IDF векторизации, чтобы преобразовать текстовые данные в числовые признаки. Другие метода векторизации текста давали либо низкий результат работы модели, либо имели недостаток в том, что модель обучалась медленно, тк вектор отзыва был очень большой и громоздкий.
- Строится модель Random Forest Classifier, которая обучается на train выборке. Другие классические алгоритмы и алгоритм BERT давали более низкий результат или проигрывали в скорости.
- Модель предсказывает класс (положительный или отрицательный отзыв) для отзывов в тестовой выборке.
- Выводится точность предсказания и classification report, который содержит метрики precision, recall, f1-score и support для каждого класса. В данном случае, точность предсказания составляет около 0.84.

3. Классификация рейтинга на основе предсказанных классов:
- Отзывы в обучающей выборке и тестовой выборке также векторизуются с помощью TF-IDF векторизации.
- К векторизованным отзывам добавляется предсказанный класс из предыдущего этапа (1 - положительный отзыв, 0 - отрицательный отзыв).
- Строится новая модель Random Forest Classifier, которая обучается на обучающей выборке с добавленным предсказанным классом.
- Модель предсказывает рейтинг для отзывов в тестовой выборке на основе предсказанных классов и текстовых признаков.
- Выводится точность предсказания для рейтингов. В данном случае, точность предсказания составляет около 0.39.

Заключение:
- Модель Random Forest Classifier показала хорошую точность при классификации тональности отзывов, но низкую точность при предсказании рейтингов на основе классифицированных отзывов.

- Причиной низкой точности предсказания рейтингов может быть сложность задачи классификации рейтинга на основе текстовых данных и наличие шума в данных, так же разделение тестовой и тренировочной выборки пополам, когда обычно соотношение иное

Применение к новым отзывам:
Обученная модель была применена для анализа введённого самостоятельно отзыва . Модель верно определила положительный тон отзыва, выделив захватывающий сюжет фильма, выдающиеся исполнения и захватывающую визуальную составляющую