Тестовое задание отчет:

1) Собрать датасет из текстовых файлов, расположенных в папках с метками pos/neg, и извлечь из названий файлов рейтинг отзыва. (в рейтинге нет оценок 5 и 6)

2) Выполнить подготовку данных, которая включает:

- избавление от пунктуации

- удаление стоп-слов (например – i, you, me

- приведение слов к начальной форме (лемматизация)

3) Получаем словарь наиболее встречающихся слов (словарь токенов)

4) Из полученных обработанных данных создаем массив определённой длины, с необходимыми словами.

5) Переводим массивы в тенсоры и создаем объект DataLoader.

6) Создаем модель на архитектуре Pytorch, используем сверточных слои, оптимизатор – Adam, критерий оценки – бинарная кроссэнтропия.

7) Итоговый Accuracy на тесте равен - 0.823.

8) Для мультиклассовой классификации используем градиентный бустинг, количество деревьев и их глубина получены с помощью кросс-валидации.

9) Итоговый Accuracy на тесте равен - 0.23084.