Отчет о выполнении тестового задания на стажировку по Data Science, Гринатом

Выполнила: Куроедова А. П.

1. Модель-основа – “bert-base-uncased”

В размеченном датасете IMDB представлены позитивные и негативные отзывы с рейтингами 1-4 и 7-10, и не представлены нейтральные (5-6).

Главной метрикой выбрана f1_score для восьми классов.

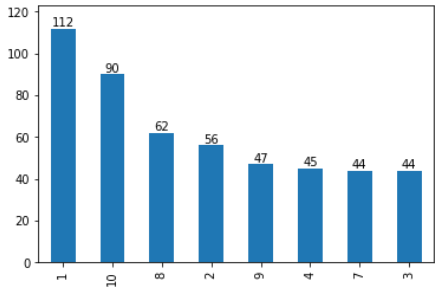
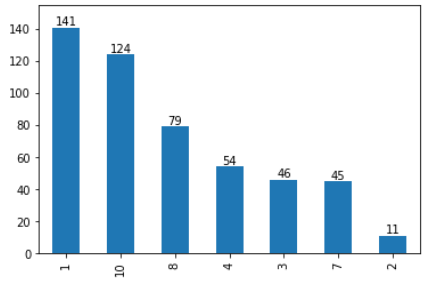
За основу была взята предобученная модель “bert-base-uncased” для задачи классификации текстов с huggingface.

Предсказание одного из восьми классов получается стандартным путем (argmax из выходного слоя). Предсказание одного из двух классов производиться на основе сравнения сумм вероятностей принадлежности классам от 1 до 4 и от 7 до 10.

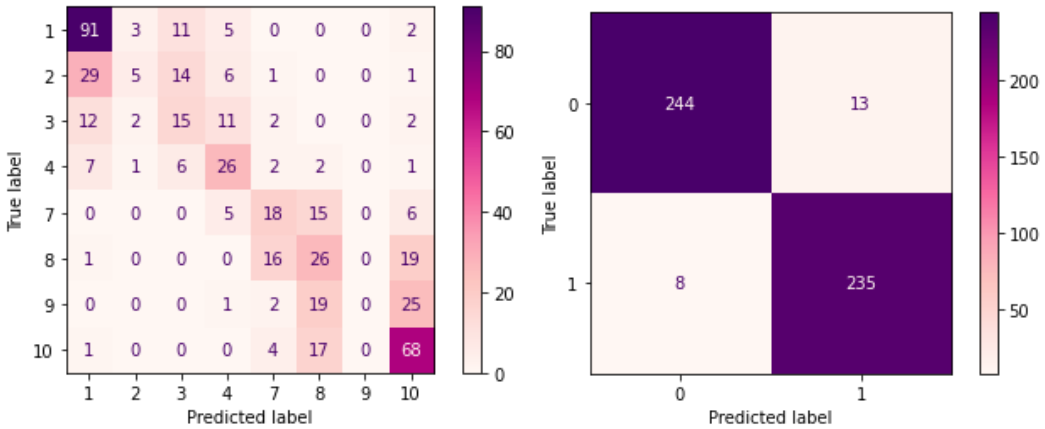
После дообучения на одной эпохе, были получены скоры:

	Для 8 классов (1-4 и 7-10)		**Для 2 классов (+/-)**	
	f1-score	**accuracy**	**f1-score**	**accuracy**
bert-base-uncased	**0.45**	**0.5**	**0.95**	**0.95**

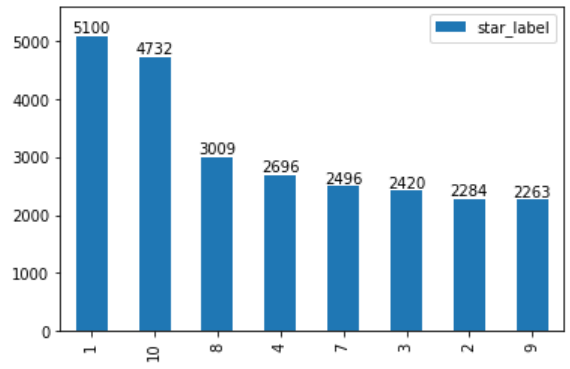
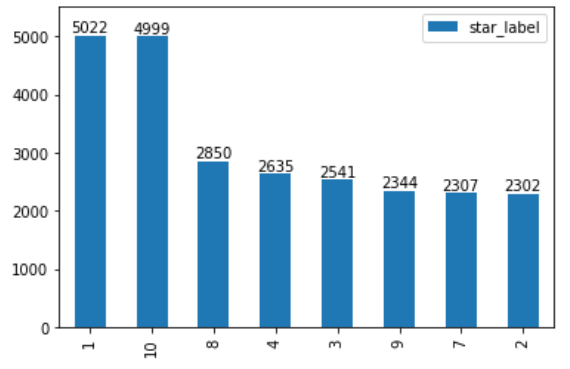
Модель недостаточно хорошо распознает классы 2 и 9, предсказав их соответственно 11 и 0 раз на фрагменте тестового датасета:

Распределение классов в тестовом датасете	Предсказания базовой модели на тестовом датасете
	

Confusion matrix:



2. Исправление дисбаланса классов

Распределение классов в тренировочном датасете	Распределение классов в тестовом датасете
	

Чтобы сгладить дисбаланс датасета, выражающийся в преобладании классов 1 и 10 с соотношением ~ 2 : 1 к каждому из остальных классов, было принято решение разделить тренировочные данные для классов 1 и 10 на половины, и каждую эпоху чередовать эти половины. Таким образом, датасет для одной эпохи будет состоять из ~2500 элементов каждого класса.

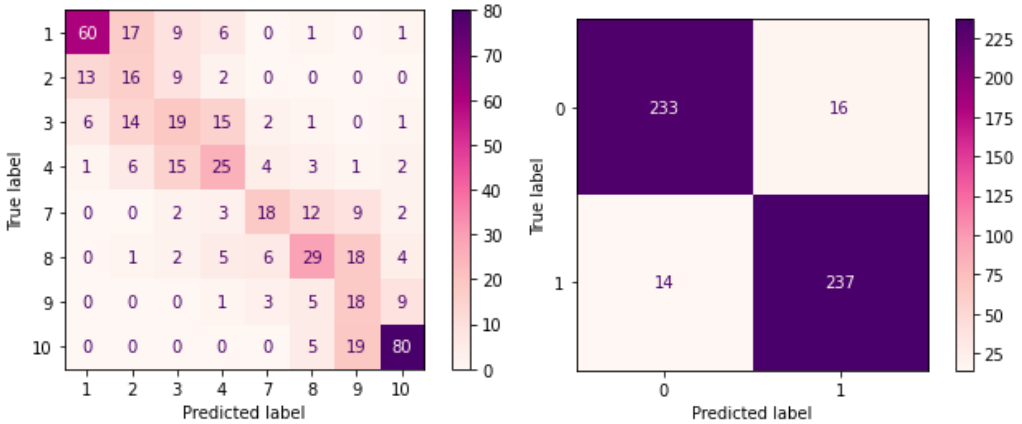
3.1. Метод предсказания на основе бинарного поиска – “bert_binary_tree”

Можно воспользоваться взаимосвязью между классами: 1-4 – отрицательно, 7-10 – положительно. 1-2 – резко отрицательно, 3-4 – слегка отрицательно, 1 - наиболее резко отрицательно, и так далее.

Было разработано дополнение базовой модели, находящее prediction по принципу бинарного дерева. Чтобы получить предсказание по вектору вероятностей, нужно сначала сравнить суммы вероятностей принадлежности к классам 1-4 и классам 7-10, а потом проделать то же самое внутри бинарных подклассов.

Оно не влияет на обучение весов модели, но позволяет делать более точные предсказания.

Результаты:



3.2. Классификация порядковых данных

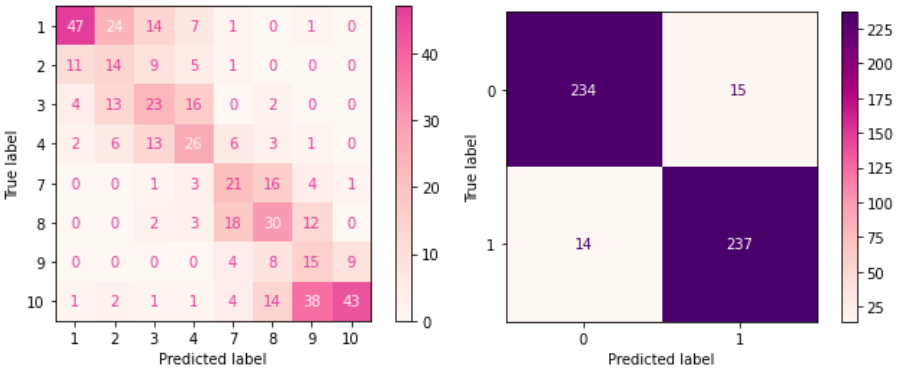
bert_multilabel_ordinary

Попробуем воспользоваться порядковостью данных. Закодируем каждый из 10 классов вектором размера 9:

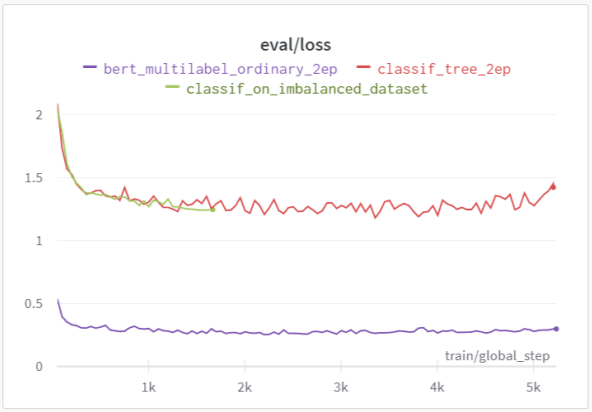
```
          >1  >2  >3  >4  >5  >6  >7  >8  >9
1   -> [0,0,0,0,0,0,0,0,0]
2   -> [1,0,0,0,0,0,0,0,0]
3   -> [1,1,0,0,0,0,0,0,0]
                  ...
8   -> [1,1,1,1,1,1,1,0,0]
9   -> [1,1,1,1,1,1,1,1,0]
10  -> [1,1,1,1,1,1,1,1,1]
```

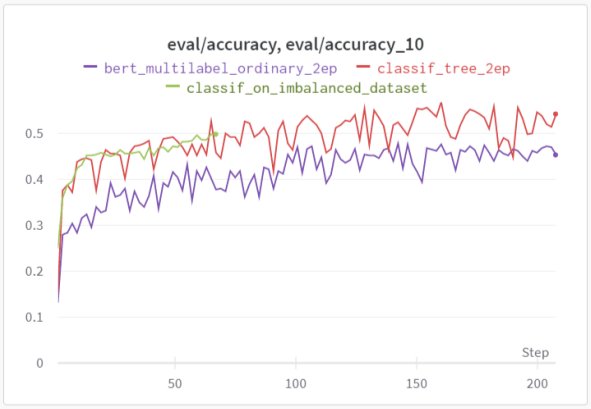
Возьмем базовую модель “bert-based-uncased”, но для задачи multilabel classification для 9 классов.

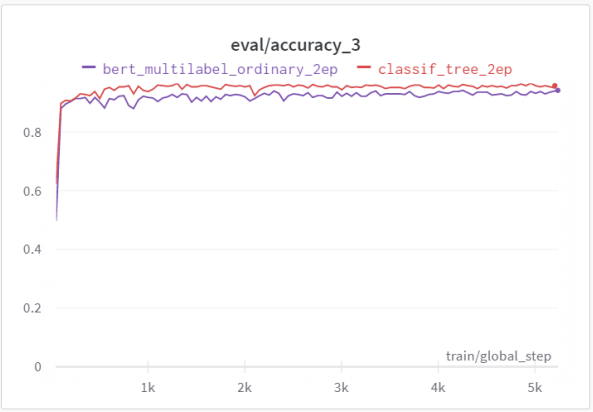
Результаты:

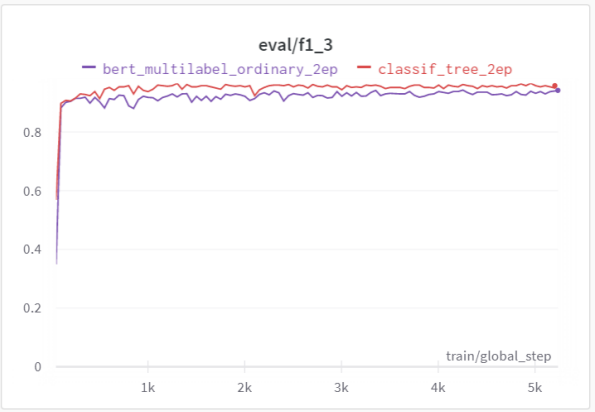


4. Графики обучения









5. Результаты на тестовой выборке

	Для 8 классов (1-4 и 7-10)		Для 2 классов (+/-)	
	f1-score	**accuracy**	**f1-score**	**accuracy**
bert_binary_tree	**0.54**	**0.53**	**0.94**	**0.94**
bert_multilabel_ordinary	**0.46**	**0.44**	**0.94**	**0.94**

6. Лучшая модель

Лучший результат показала модель “bert_binary_tree”.

7. Нейтральный класс

Для инференса этой модели установим порог 0.2. Если модуль разности вероятностей принадлежности отзыва к положительному и негативному классу меньше порога, то считаем, что он принадлежит нейтральному классу.