Отчет

Оглавление

Введение

Была поставлена задача:

1. Обучить модель на языке Python для классификации отзывов.

2. Разработать веб-сервис на базе фреймворка Django для ввода отзыва о фильме с автоматическим присвоением рейтинга (от 1 до 10) и статуса комментария (положительный или отрицательный).

Разделим обучение на две части:

- Предсказание статуса комментария (положительный или отрицательный), другими словами задача классификации.

- Предсказание оценки оставленный комментатором будет решаться с использованием модели регрессии, так как функционал этих моделей больше подходит под эту задачу чем моделей классификации.

Классификация

Векторное представление текста

После приведения к начальным формам и удаления стоп слов были использованы 3 различных способа векторизации текста:

- Bag of Word
- TF-IDF
- Word2Vec

Выбор модели

Модели использованные для задачи классификации:

Logistic Regression, Linear SVM, RBF SVM, Random Forest, LightGBM, CatBoost, BERT

Ниже в таблице представлены результаты каждой из моделей на тесте и валидации.

Model	Validation accuracy	Test accuracy
Bag of Word + Logistic Regression	87,65	85,61
TF-IDF + Logistic Regression	87,16	86,80
Word2Vec + Logistic Regression	87,36	87,18
Bag of Word + Linear SVM	85,38	83,53
TF-IDF + Linear SVM	89,10	88,26
Word2Vec + Linear SVM	79,78	78,90
Bag of Word + RBF SVM	86,69	86,04
TF-IDF + RBF SVM	88,73	87,48
Word2Vec + RBF SVM	88,04	87,43
Bag of Word + Random Forest	85,50	84,70
TF-IDF + Random Forest	84,61	84,60
Word2Vec + Random Forest	83,22	82,60
Bag of Word + LightGBM	86,06	85,36
TF-IDF + LightGBM	84,86	85,08
Word2Vec + LightGBM	85,53	85,32
CatBoost	90,57	88,39
CatBoost + top 300 TF-IDF	90,26	88,70
BERT	89,06	89,57

Зеленым выделены лучшие результаты, на validation это CatBoost, на test-е BERT

Так же синим выделены интересные результаты: TF-IDF + Linear SVM и CatBoost + top 300 TF-IDF, первая модель не смотря на свою простату показала максимально близкий к бустингам и BERT-у результат.
Отдельно стоит уделить больше внимания отбору фич из TF-IDF по критерию хи-квадрат.

worst, waste, awful,
great, wonderful, terrible,
waste time, boring, horrible
stupid, worse, wonderful
excellent, crap, poor
lame, worst movie, nothing,
love, poorly, minutes
avoid, pointless, perfect
best, ridiculous

Из top25 фич можно сделать выводы о том, какие слова или выражения являются наиболее значимыми для классификации. Например, "**worst**" (худший), "**awful**" (ужасный), "**great**"

(великолепный), "**terrible**" (ужасный) "**waste time**" (тратить время) "**boring**" (скучный) оказываются статистически важными признаками, что соответствует реальности.

Ещё интересным кажется, что модели работавшие с Word2Vec, казалось бы более новым методом векторизации, показывают худший результат. Это можно объяснить тем, что используемый Word2Vec был обучен на сильно меньшем корпусе, в отличии от аналогичной модели Google, однако последняя работает слишком долго.

Вывод

В сводном анализе результатов оказалось, что BERT продемонстрировал лучшие показатели. Однако, при принятии решения о выборе модели для MVP, рекомендуется использовать CatBoost из-за его относительной простоты и незначительного отставания в точности от BERT менее чем на 1%.

Регрессия

Покажем, почему задача регрессии лучше подходит для предсказания оценок.

В случае регрессии loss функция выглядит следующим образом (для примера MAE)

$$Loss = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i|$$

А в задаче классификации так(для примера -Энтропия):

$$Loss = \sum_{i=0}^{N} p_i \log p_i$$

В таком случае для задачи регрессии гораздо принципиальней когда $y_{pred} = 9$ а $y_{true} = 1$, и на оборот менее принципиально, когда $y_{pred} = 5$ а $y_{true} = 4$, для энтропии Loss одинаково изменится для этих двух ситуаций. Нам в задаче важнее максимально приблизиться к ответу нежели точно в него попасть. Из этого можно сделать вывод, что регрессия будет лучшим решение для предсказания рейтинга, поставленного пользователем.

Выбор модели

Для данной задачи использовались следующие модели:

Linear Regression, LightGBM, CatBoost

Результаты приведены в таблице ниже

Model	Validation MAE	Test MAE
Bag of Word + Linear Regression	2,63	2,73
TF-IDF + Linear Regression	1,66	1,61
Word2Vec + Linear Regression	2,02	2,09
Bag of Word + LightGBM	1,79	1,81
TF-IDF + LightGBM	1,76	1,76
Word2Vec + LightGBM	1,73	1,77
CatBoost	1,65	1,66

Вывод

Задача предсказания оценки кажется более сложной чем задача классификации, так как один и тот же отзыв от разных людей может иметь разные оценки.

Финальной моделью опять был выбран CatBoost.

Заключение

- Было проведено исследование каждой из частей задания
- Найдено оптимальное решение для MVP
- Создан web-прототип на основе Streamlit