Построение модели классификации отзывов. Разработка веб-сервиса на Django.

Исходные данные: https://ai.stanford.edu/~amaas/data/sentiment/

открытый набор данных «Large Movie Review Dataset v1.0» от Imdb, который содержит в себе отзывы о фильмах, а также соответствующие им оценки рейтинга.

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).

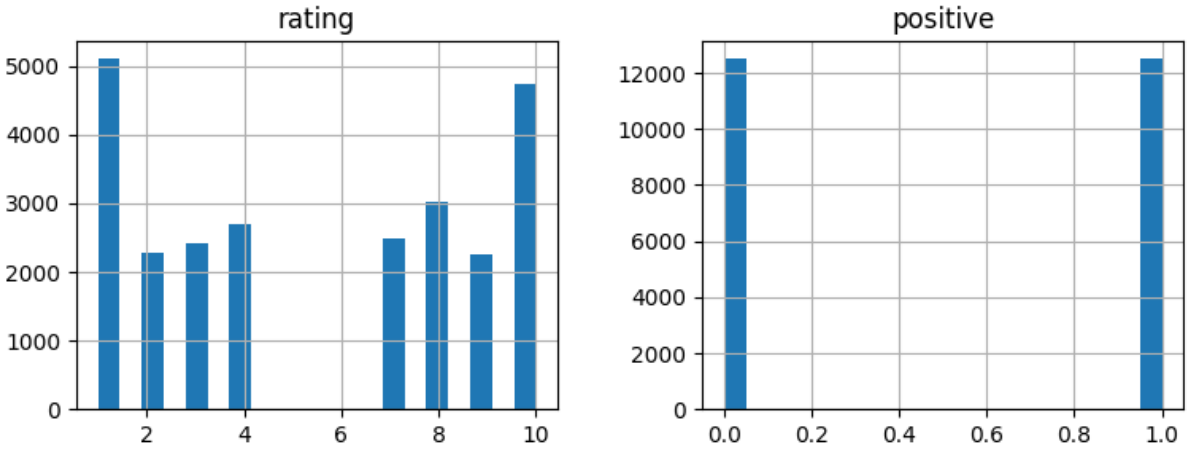
Отчет о проделанной работе

В рамках исследования были проведены следующие этапы:

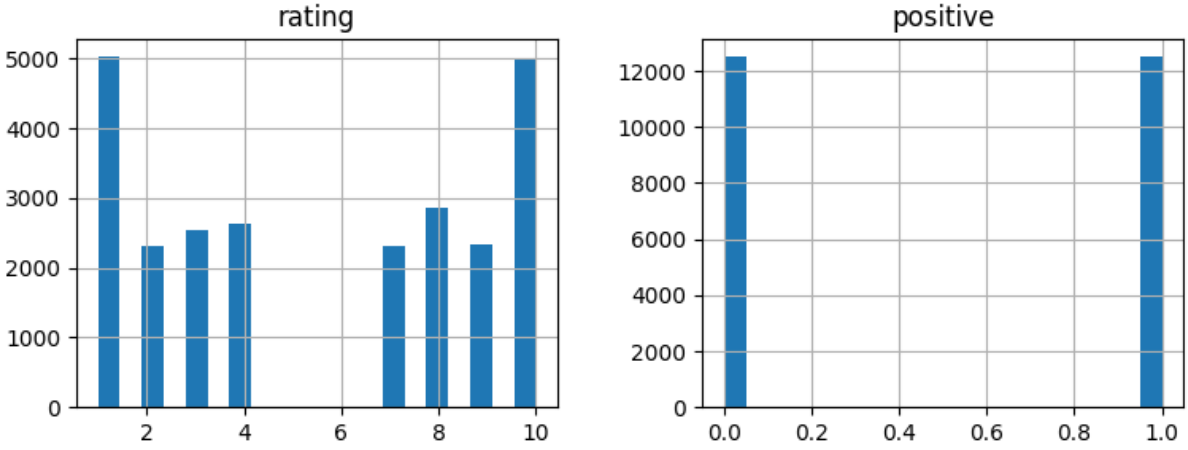
1. Загрузка и изучение данных

1.1. Обучающая и тестовая выборки содержат по 25 тыс. записей с отзывами о фильмах, оценками (1-4 и 7-10) и статусами (1 - позитивный, 0 - негативный).
1.2. Отзывы с нейтральным рейтингом 5-6 не включены в обучающий и тестовый наборы данных.
1.3. Пропусков в данных не обнаружено.
1.4. Классы статуса сбалансированы 1:1 в обеих выборках (по 12,5 тыс.).
1.5. Классы оценок в обучающей и тестовой выборках имеют похожее распределение: самые популярные оценки - 1 и 10 (по ~5 тыс.); остальные оценки встречаются реже, но распределены по классам достаточно равномерно (от ~2 до 3 тыс.).

Распределение классов в обучающем наборе данных:



Распределение классов в тестовом наборе данных:



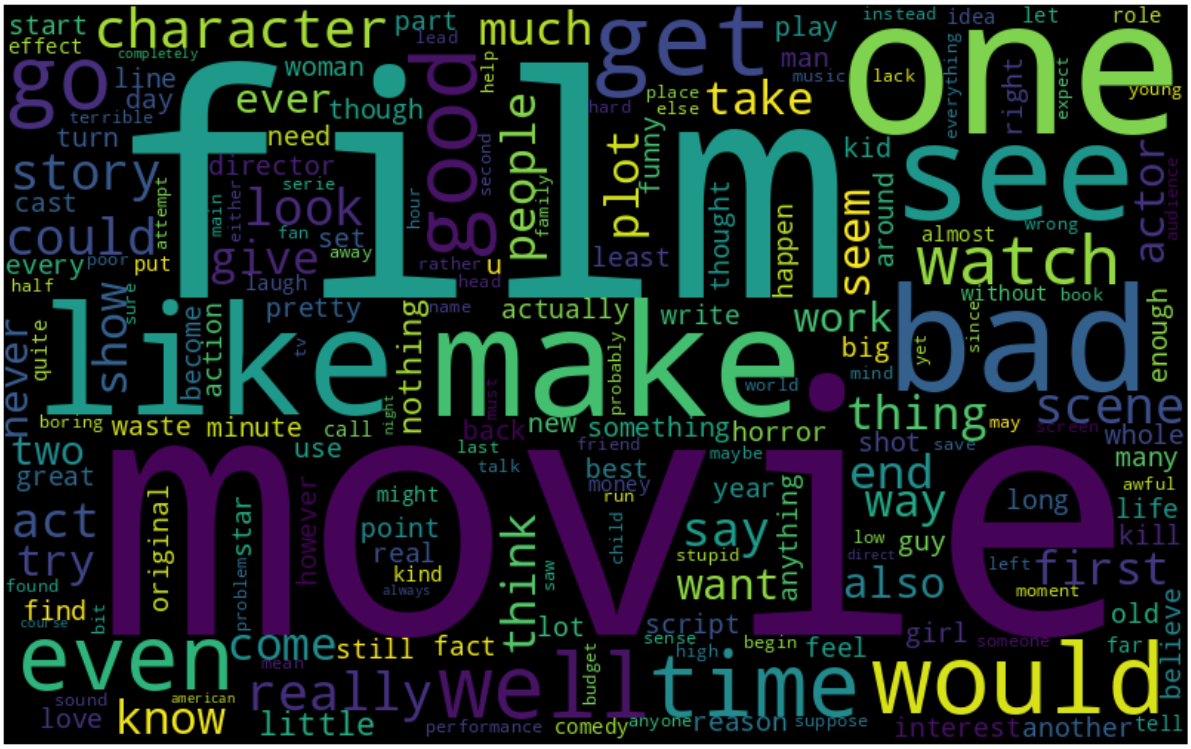
2. Обработка текста

2.1. Перед обучением моделей комментарии были очищены от лишних символов, спецсимволов и стоп-слов, проведена лемматизация с помощью WordNetLemmatizer().

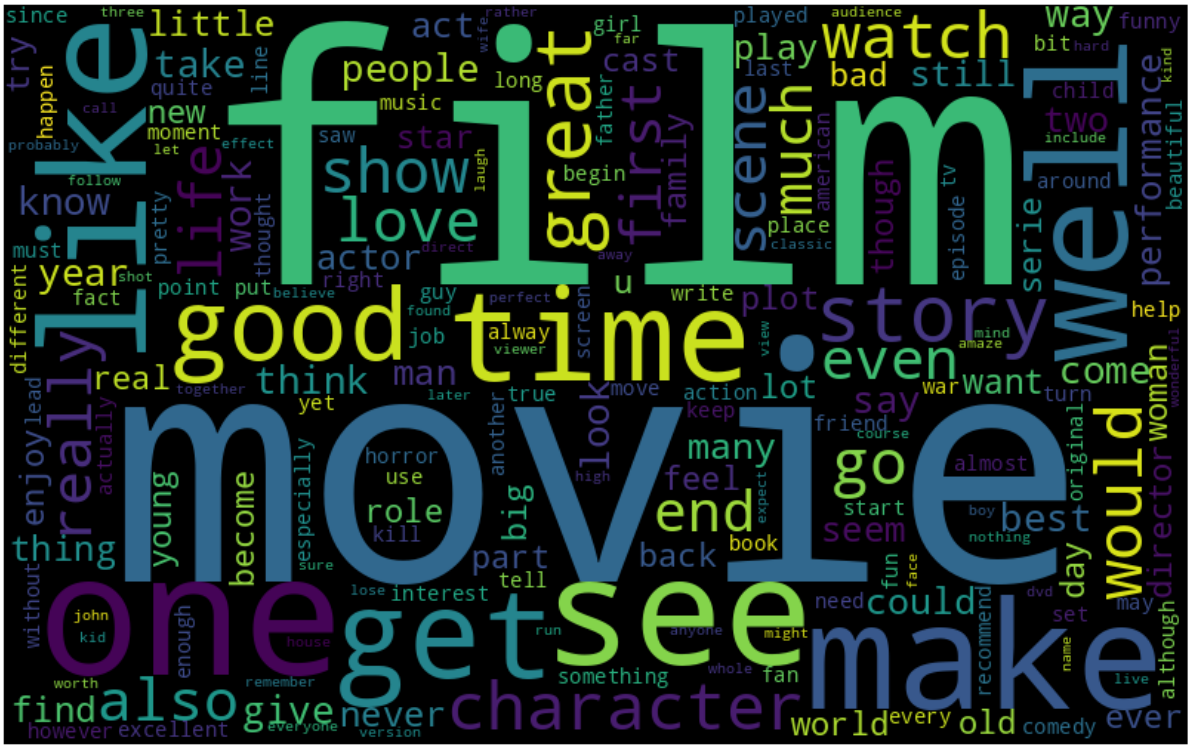
2.2. После обработки из обучающего набора были обнаружены и удалены 100 дубликатов отзывов.

2.3. Вывели облака слов для негативных и позитивных отзывов: наиболее популярные слова во многом совпадают (film, movie, one, see, make, like и др.), но при этом есть и заметные отличия: например, в негативных отзывах намного чаще встречается слово bad, а в позитивных - слово love.

Most frequently used words in Negative reviews



Most frequently used words in Positive reviews



3. Предсказание тональности отзывов (positive / negative)

- Перед подачей текста в модель будем проводить его TF-IDF преобразование с помощью TfidfVectorizer().
- Так как классы целевого признака сбалансированы, то качество модели можно оценивать метрикой accuracy (доля верных ответов).

3.1. Классические ML-модели

Обучение, подбор параметров и выбор лучшей модели проводили на кросс-валидации с помощью RandomizedSearchCV для моделей:

- LogisticRegression
- SGDClassifier
- LGBMClassifier
- DummyClassifier (для оценки адекватности моделей)

Результаты лучших-моделей на кросс-валидации:

	model	accuracy	roc_auc	recall	precision	params
0	SGDClassifier	0.891	0.891	0.904	0.881	{'sgdclassifier__loss': 'hinge', 'sgdclassifie...
1	LogisticRegression	0.888	0.888	0.898	0.881	{'logisticregression__solver': 'saga', 'logist...
2	LGBMClassifier	0.870	0.870	0.877	0.866	{'lgbmclassifier__n_estimators': 300, 'lgbmcla...
3	DummyClassifier	0.501	0.501	1.000	0.501	{'dummyclassifier__strategy': 'most_frequent'}

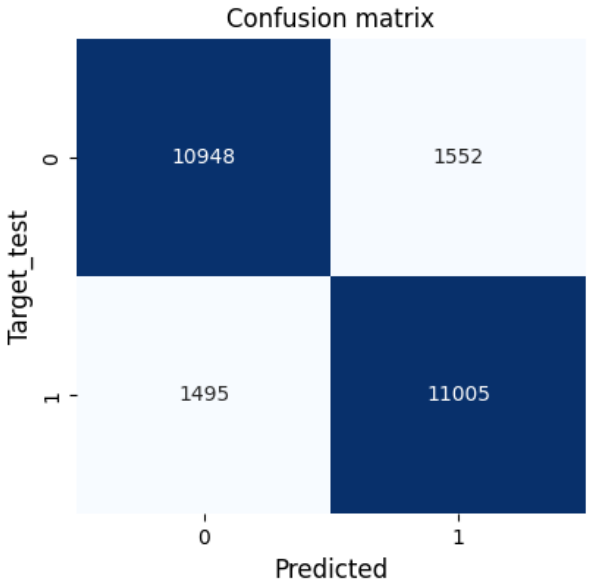
Лучшие результаты на кросс-валидации (accuracy 89,1%) показала модель SGDClassifier с параметрами:

{'loss': 'hinge', 'alpha': 0.0001}.

Результаты лучшей модели **на тестовых данных**:

	precision	recall	f1-score	support
0	0.88	0.88	0.88	12500
1	0.88	0.88	0.88	12500
accuracy			0.88	25000
macro avg	0.88	0.88	0.88	25000
weighted avg	0.88	0.88	0.88	25000

На тестовой выборке **accuracy = 88%**, что является неплохим результатом. Попадание в классы равномерное:



3.2. Нейросеть

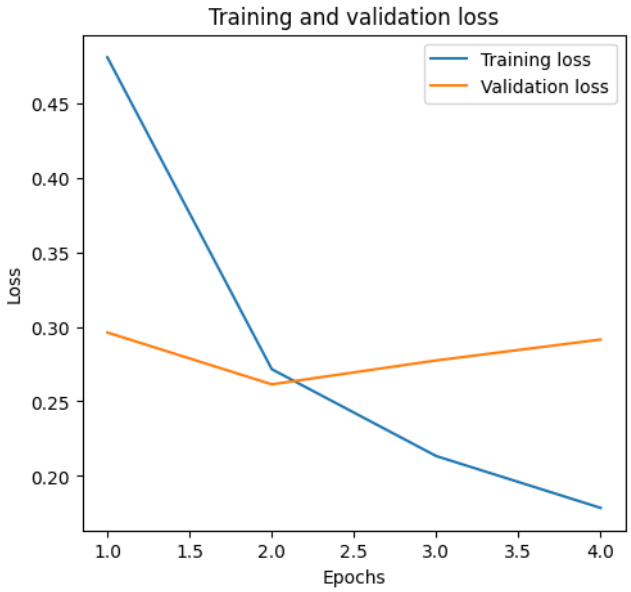
Построили простую нейросеть для бинарной классификации:

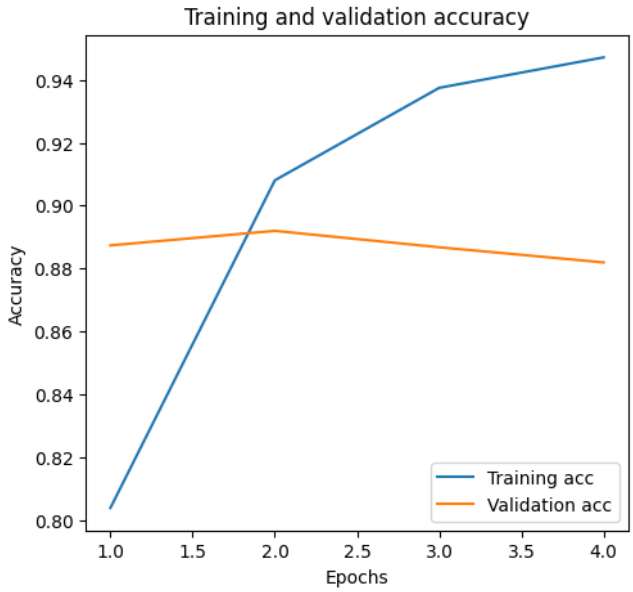
```
def build_neural_network():
    model = tf.keras.Sequential([
        layers.Dense(64, input_shape=(10000, )),
        layers.Dropout(0.5),
        layers.Dense(10, activation='relu'),
        layers.Dropout(0.5),
        layers.Dense(1, activation='sigmoid')

])
    model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['binary_accuracy'])

    return model
```

Результаты обучения:

```
Epoch 1/4
200/200    - loss: 0.4807 - binary_accuracy: 0.8039 - val_loss: 0.2962 - val_binary_accuracy: 0.8873
Epoch 2/4
200/200    - loss: 0.2716 - binary_accuracy: 0.9080 - val_loss: 0.2614 - val_binary_accuracy: 0.8920
Epoch 3/4
200/200    - loss: 0.2134 - binary_accuracy: 0.9374 - val_loss: 0.2775 - val_binary_accuracy: 0.8867
Epoch 4/4
200/200    - loss: 0.1786 - binary_accuracy: 0.9471 - val_loss: 0.2915 - val_binary_accuracy: 0.8819
```

Для наглядности построили графики изменения точности и потерь в зависимости от количества эпох. Видим, что на валидации потери после 2 эпох начинают возрастать, а точность начинает падать. Это говорит о переобучении модели после 2 эпох:





Переобучили нейросеть на 2 эпохах:

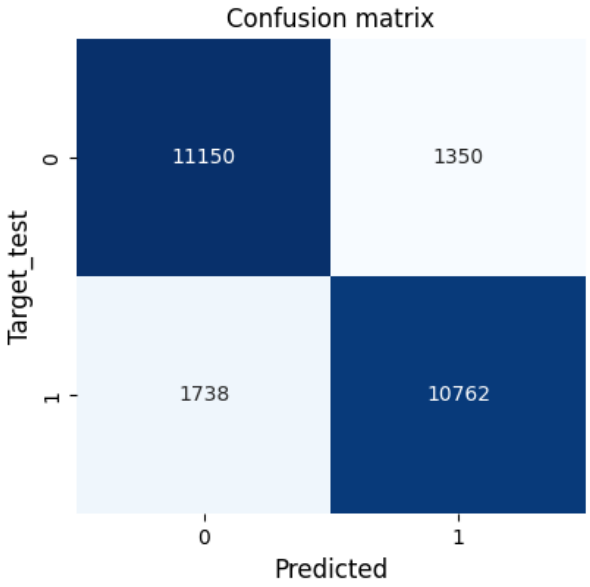
```
Epoch 1/2
200/200    - loss: 0.5149 - binary_accuracy: 0.7787 - val_loss: 0.3115 - val_binary_accuracy: 0.8809
Epoch 2/2
200/200    - loss: 0.2806 - binary_accuracy: 0.9055 - val_loss: 0.2616 - val_binary_accuracy: 0.8908
```

На валидации accuracy = 89,1%. Это результат сопоставимый с выбранной лучшей ML-моделью SGDClassifier. Вероятно, что более высокую точность можно достичь с помощью эмбиддингов на предобученной модели BERT, но данный процесс занимает много времени и требует более высокие вычислительные ресурсы.

Результаты нейросети **на тестовых данных**:

	precision	recall	f1-score	support
0	0.85	0.90	0.88	12500
1	0.90	0.84	0.87	12500
accuracy			0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

Видим, что результат даже немного хуже, чем у SGDClassifier, и предсказание классов менее равномерное:



Вывод: для предсказания тональности отзывов остановим свой выбор на более простой модели SGDClassifier, которая дает достаточно хорошие результаты (accuracy = 88%) с равномерным попаданием в классы.

4. Предсказание рейтинга

- Перед подачей текста в модель будем проводить его TF-IDF преобразование с помощью TfidfVectorizer().
- Так как классы целевого признака несбалансированны, то качество модели можно оценивать средневзвешенным по классам accuracy.

4.1. Классические ML-модели

LGBMClassifier – результаты на валидации:

	precision	recall	f1-score	support
1	0.57	0.64	0.61	1013
2	0.22	0.16	0.18	454
3	0.20	0.17	0.19	483
4	0.29	0.29	0.29	536
7	0.29	0.26	0.27	499
8	0.28	0.28	0.28	601
9	0.18	0.14	0.16	451
10	0.49	0.59	0.54	943
accuracy			0.38	4980
macro avg	0.31	0.32	0.31	4980
weighted avg	0.36	0.38	0.37	4980

SGDClassifier – результаты на валидации:

	precision	recall	f1-score	support
1	0.56	0.72	0.63	1013
2	0.18	0.12	0.15	454
3	0.22	0.18	0.19	483
4	0.29	0.30	0.30	536
7	0.27	0.28	0.28	499
8	0.27	0.22	0.24	601
9	0.22	0.18	0.20	451
10	0.53	0.59	0.56	943
accuracy			0.39	4980
macro avg	0.32	0.32	0.32	4980
weighted avg	0.36	0.39	0.37	4980

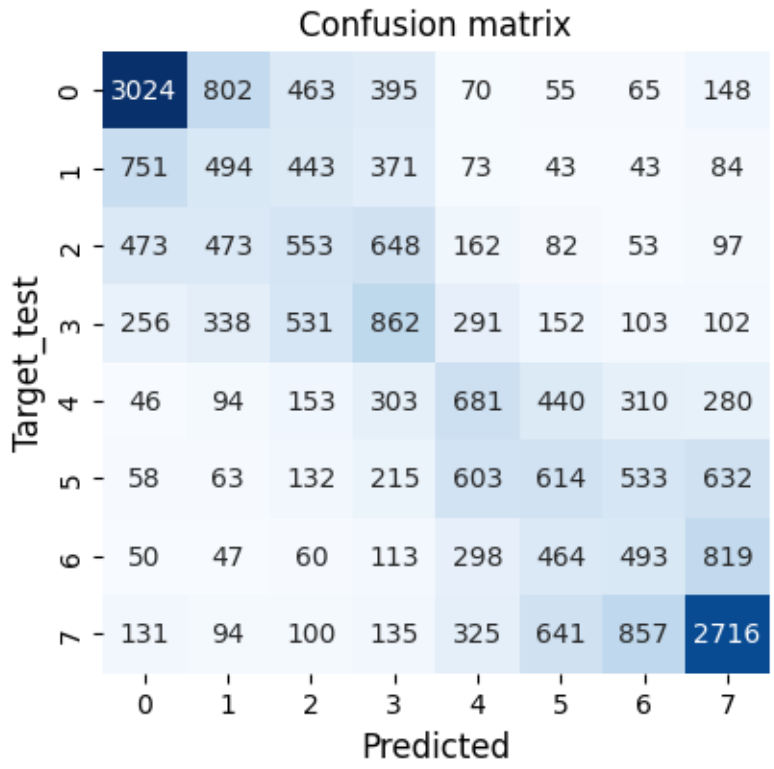
LogisticRegression – результаты на валидации:

	precision	recall	f1-score	support
1	0.63	0.57	0.60	1013
2	0.24	0.23	0.24	454
3	0.24	0.25	0.24	483
4	0.31	0.33	0.32	536
7	0.31	0.36	0.33	499
8	0.29	0.27	0.28	601
9	0.21	0.24	0.22	451
10	0.55	0.54	0.54	943
accuracy			0.39	4980
macro avg	0.35	0.35	0.35	4980
weighted avg	0.40	0.39	0.39	4980

Лучшие результаты на валидации показала LogisticRegression с параметрами {class_weight: 'balanced', max_iter: 300}

Результаты лучшей модели **на тестовых данных**:

	precision	recall	f1-score	support
1	0.63	0.60	0.62	5022
2	0.21	0.21	0.21	2302
3	0.23	0.22	0.22	2541
4	0.28	0.33	0.30	2635
7	0.27	0.30	0.28	2307
8	0.25	0.22	0.23	2850
9	0.20	0.21	0.21	2344
10	0.56	0.54	0.55	4999
accuracy			0.38	25000
macro avg	0.33	0.33	0.33	25000
weighted avg	0.38	0.38	0.38	25000



Результаты именно точных ответов не очень высокие (**accuracy = 38%**), что объясняется сложностью точного предсказания мультиклассов по сравнению с бинарной классификацией. Однако по матрице ошибок видно достаточно неплохое попадание в соседние (близкие) классы.

4.2. Нейросеть

Построили простую нейросеть для множественной классификации:

```
def build_neural_network_rat():
    model = tf.keras.Sequential([
        layers.Dense(128, input_shape=(10000, )),
        layers.Dropout(0.5),
        layers.Dense(32, activation='relu'),
        layers.Dropout(0.5),
        layers.Dense(11, activation='softmax')

])
    model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['sparse_categorical_accuracy'])

    return model
```

```
Epoch 1/4
200/200    - loss: 1.9728 - sparse_categorical_accuracy: 0.2918 - val_loss: 1.6043 - val_sparse_categorical_accuracy: 0.3882
Epoch 2/4
200/200    - loss: 1.5356 - sparse_categorical_accuracy: 0.4200 - val_loss: 1.4624 - val_sparse_categorical_accuracy: 0.4299
Epoch 3/4
200/200    - loss: 1.3735 - sparse_categorical_accuracy: 0.4699 - val_loss: 1.4580 - val_sparse_categorical_accuracy: 0.4245
Epoch 4/4
200/200    - loss: 1.2766 - sparse_categorical_accuracy: 0.5051 - val_loss: 1.4877 - val_sparse_categorical_accuracy: 0.4237
```

Результаты обучения:

	precision	recall	f1-score	support
1	0.53	0.81	0.64	1013
2	0.00	0.00	0.00	454
3	0.33	0.10	0.15	483
4	0.29	0.45	0.35	536
7	0.29	0.23	0.26	499
8	0.31	0.32	0.31	601
9	0.50	0.00	0.01	451
10	0.48	0.74	0.58	943
accuracy			0.42	4980
macro avg	0.34	0.33	0.29	4980
weighted avg	0.37	0.42	0.36	4980

Вывод: несмотря на то, что на валидации общая метрика accuracy = 42% у нейросети получилась больше, чем у LogisticRegression, взвешенный показатель по классам при этом хуже (36% vs 39%): классы 2 и 9 вообще имеют нулевое f1-score. Вероятно, что более высокую точность можно достичь с помощью усложнения модели, например эмбиддингов на предобученной модели BERT, но в условиях ограниченности ресурсов остановим свой выбор на простой модели LogisticRegression.

5. Разработка веб-сервиса на Django

Для тестирования:

http://supermi3.beget.tech/

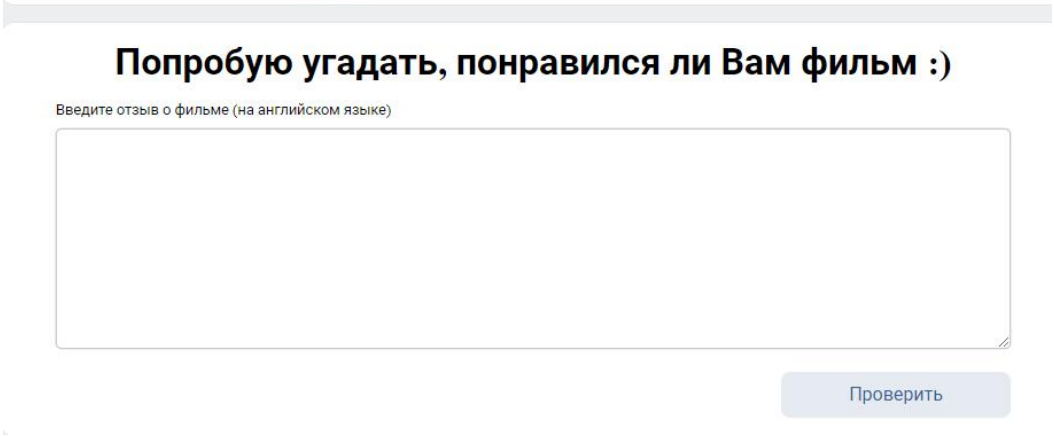
Проект расположен в репозитории по ссылке:

https://github.com/ElenaWF/greenatom

Для запуска приложения на Windows локально скачайте репозиторий и в папке выполните команды:

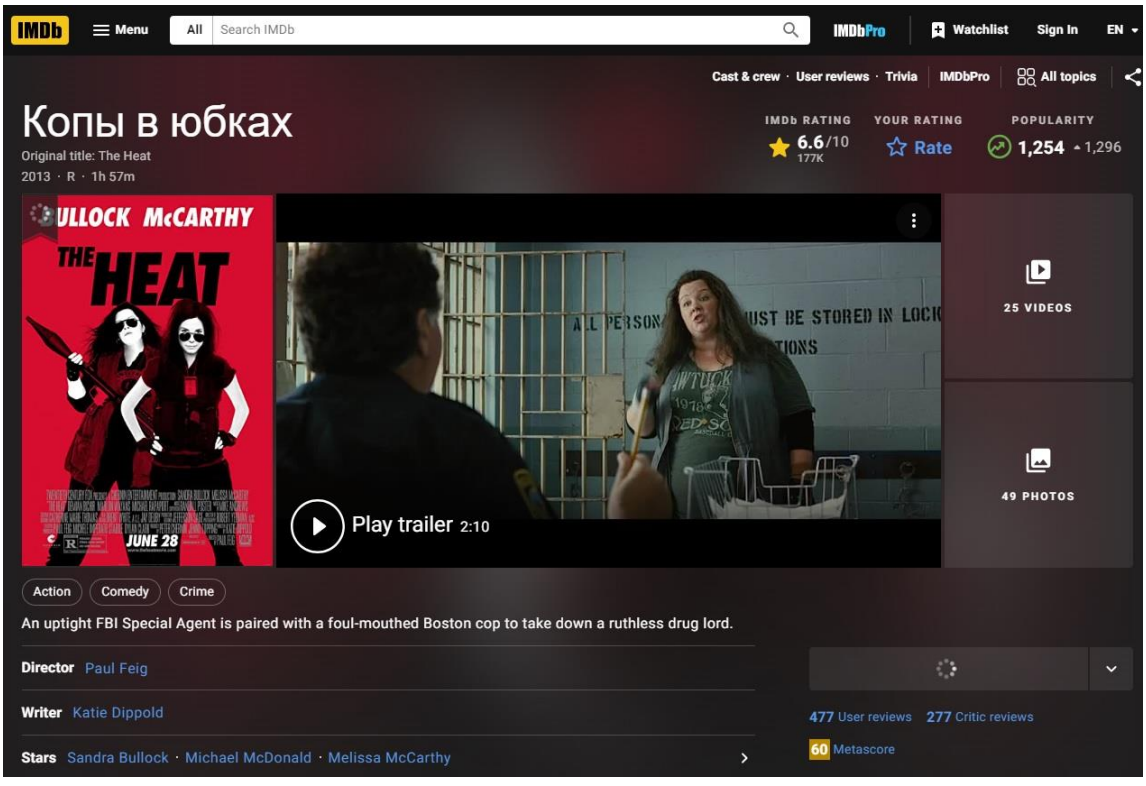
1. py -m venv env
2. .\env\Scripts\activate
3. pip install -r requirements.txt
4. py manage.py runserver

Проект будет доступен локально http://127.0.0.1:8000/



5. Тестирование сервиса

Возьмем реальные отзывы с сайта IMDb на фильм



2/10

Appallingly bad
kingbad 30 June 2013

Warning: Spoilers

I have no problem with remakes. Considering how much it costs to make a movie these days, selling something that people enjoyed the first time makes sense; it's a predictable return on the studio's investment, and it's easy to market a movie that's just like "so-and-so meets such-and- such". So, when a studio wants to remake an 80s buddy-cop comedy, with women in the lead roles, I say why not? There are enough Murphy-Nolte, Gibson-Glover, and Willis-everybody role models out there, how could you go wrong?

Here's how. Instead of ripping off a successful buddy-cop franchise, The Heat has Running Scared meeting The Odd Couple. Instead of a completely unconvincing pairing of a Borscht Belt ham and a tap dancer, they've got Sandra Bullock playing, yet again, a neurotic tight-ass and Melissa McCarty playing, yet again, a foul-mouthed slob. These two completely unlikable, and unbelievable, characters, one an FBI agent who somehow manages to confound detection dogs with her ability to find hidden drugs and weapons despite a complete lack of detection skills, and the other a Boston cop who (somehow) manages to successfully work undercover, in her own neighborhood, despite being morbidly obese, obnoxiously loud, and wearing the same clothes for days at a time. Both are (deservedly) pariahs within their respective departments, who somehow succeed despite their own individual and collective incompetence, and grow to adopt each other's most annoying bad habits. This movie plays to an audience's lowest, dumbest instincts- sadly, it will probably be a hit. Avoid at all costs.

166 out of 343 found this helpful. Was this review helpful? Sign in to vote.

Permalink

Результат анализа отзыва

I have no problem with remakes. Considering how much it costs to make a movie these days, selling something that people enjoyed the first time makes sense; it's a predictable return on the studio's investment, and it's easy to market a movie that's just like "so-and-so meets such-and- such". So, when a studio wants to remake an 80s buddy-cop comedy, with women in the lead roles, I say why not? There are enough Murphy-Nolte, Gibson-Glover, and Willis-everybody role models out there, how could you go wrong?

Here's how. Instead of ripping off a successful buddy-cop franchise, The Heat has Running Scared meeting The Odd Couple. Instead of a completely unconvincing pairing of a Borscht Belt ham and a tap dancer, they've got Sandra Bullock playing, yet again, a neurotic tight-ass and Melissa McCarty playing, yet again, a foul-mouthed slob. These two completely unlikable, and unbelievable, characters, one an FBI agent who somehow manages to

Отзыв: Негативный
Рейтинг: 2

Попробовать снова

Ура, верно!

7/10

A fun buddy movie with two female law enforcement officers

Tweekums 16 January 2017

Warning: Spoilers

Sarah Ashburn is a by-the-book FBI agent who is keen for promotion; unfortunately her colleagues find her arrogant. Shannon Mullins is a foul mouthed Boston cop who is happy to beat a confession of a suspect and scares her colleagues even more than the criminals. They make unlikely partners but when Ashburn is sent to Boston it identify and arrest drug lord 'Mr. Larkin' they are forced to work together. At first they don't get along but inevitably they end up a fine team as they move closer to Larkin; go through various dangers and learn more about each other.

There are lots of buddy movies featuring mismatched male cops but I think this first distaff take on the genre I've seen. The story is fairly typical of the genre but there is still a bit of a twist concerning Larkin's identity. Sometimes this type of film is mostly action with a few comedy moments and other times the comedy is the main selling point□ this is definitely in the latter camp. Melissa McCarthy provides most of the laughs as Mullin, although some viewers will be put off by the character's constant foul language. Sandra Bullock also does a fine job as Ashburn even though that means playing the straight-guy most of the time. The two develop a fine chemistry as the film progresses. There is a decent amount of action and one particularly wince inducing moment when Ashburn gets stabbed in the leg. Overall I'd recommend this to anybody wanting a good laugh so long as they aren't easily offended.

9 out of 12 found this helpful. Was this review helpful? Sign in to vote.

Permalink

Результат анализа отзыва

Sarah Ashburn is a by-the-book FBI agent who is keen for promotion; unfortunately her colleagues find her arrogant. Shannon Mullins is a foul mouthed Boston cop who is happy to beat a confession of a suspect and scares her colleagues even more than the criminals. They make unlikely partners but when Ashburn is sent to Boston it identify and arrest drug lord 'Mr. Larkin' they are forced to work together. At first they don't get along but inevitably they end up a fine team as they move closer to Larkin; go through various dangers and learn more about each other.

There are lots of buddy movies featuring mismatched male cops but I think this first distaff take on the genre I've seen. The story is fairly typical of the genre but there is still a bit of a twist concerning Larkin's identity. Sometimes this type of film is mostly action with a few comedy moments and other times the comedy is the

Отзыв: Позитивный

Рейтинг: 7

Попробовать снова

И снова верно!

10/10

Bullock and McCarthy Bring "The Heat"

jon.h.ochiai 4 July 2013

"The Heat" is hysterical. Sandra Bullock and Melissa McCarthy are awesome! "The Heat" is the funniest movie of the year. I laughed out loud a lot. Yes, Sandra and Melissa reliably play in position. Bullock is Sarah Ashburn, the rigid; know it all, FBI agent booking for an Agency promotion. But as current boss Hale (patient and dashing Demian Bichir) explains her downside, "Nobody likes you." Melissa McCarthy is abrasive, no nonsense Boston Cop Mullins, who can beat the crap out of any man. She torments her Captain Woods (prematurely aged funny Tom Wilson), looking for his "lady balls" in his office.

While wallowing in their sorrows at Mullin's favorite bar, Ashburn (Bullock) confesses to Mullins (McCarthy) that not a lot of people know that she was married. With Scotch in hand, Mullins asks, "Was he a hearing man?" Director Paul Feig ("Bridesmaids") is genius with niche R-rated comedy starring women, and is blessed with Bullock and McCarthy's A-Games. Writer Katie Dippold (of "Parks and Recreation") is brilliant given a very predictable movie scenario. Will Ashburn and Mullins become BFFs? Of course. Dippold's comic Zen lies in the journey. "The Heat" is more than just "Lethal Weapon" meets "The Hangover". There is a signature moment in diner where Bullock attempts to save a choking man. Feig is comically ruthless. Bullock and McCarthy never waver out of character as their partnership naturally evolves—they are amazing.

Результат анализа отзыва

"The Heat" is hysterical. Sandra Bullock and Melissa McCarthy are awesome! "The Heat" is the funniest movie of the year. I laughed out loud a lot. Yes, Sandra and Melissa reliably play in position. Bullock is Sarah Ashburn, the rigid; know it all, FBI agent booking for an Agency promotion. But as current boss Hale (patient and dashing Demian Bichir) explains her downside, "Nobody likes you." Melissa McCarthy is abrasive, no nonsense Boston Cop Mullins, who can beat the crap out of any man. She torments her Captain Woods (prematurely aged funny Tom Wilson), looking for his "lady balls" in his office.

While wallowing in their sorrows at Mullin's favorite bar, Ashburn (Bullock) confesses to Mullins (McCarthy) that not a lot of people know that she was married. With Scotch in hand, Mullins asks, "Was he a hearing man?" Director Paul Feig ("Bridesmaids") is genius with niche R-rated comedy starring women, and is blessed with

Отзыв: Позитивный

Рейтинг: 10

Попробовать снова

В десяточку!