Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-
МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики
Бакалаврська програма

Розробка рекомендаційної системи фільмів на основі психологічного стану користувача

Текстова частина до курсової роботи
за спеціальністю «Інженерія програмного забезпечення» - 121

Керівник курсової роботи
Борозенний С.О.
_____
(Підпис)
“___” _____ 2025 року
Виконала студентка БП ІПЗ-3
Третяк Є.М.
“___” _____ 2025 року

Київ 2025

Київ 2025

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри мультимедійних систем,

доц., канд.наук

О.П. Жежерун

(підпис)

“___” _____ 2025 року

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студентці Третяк Єлизаветі Максимівні факультету інформатики 3 курсу

ТЕМА: РОЗРОБКА РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ФІЛЬМІВ НА ОСНОВІ ПСИХОЛОГІЧНОГО СТАНУ КОРИСТУВАЧА

Зміст ТЧ до курсової роботи:

Індивідуальне завдання

Анотація

Вступ

Огляд наявних рішень

Збір та підготовка даних

Навчання та застосування моделі

Генерування рекомендацій

Висновки

Список літератури

Тема: **Розробка рекомендаційної системи фільмів на основі психологічного стану користувача**

Календарний план виконання роботи:

№	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу	09.10.2024	
2.	Огляд наявних схожих розробок	22.01.2025	
3.	Аналіз технічної літератури за темою роботи	20.02.2025	
4.	Пошук та обробка даних для їх пізнішого використання	25.02.2025	
5.	Розробка рекомендаційного вебзастосунку	30.04.2025	
6.	Написання текстової частини роботи	30.04.2025	
7.	Створення слайдів для доповіді та написання доповіді	05.05.2025	
8.	Аналіз отриманих результатів із керівником	05.05.2025	
9.	Коригування роботи за результатами перевірки керівником	05.05.2025	
10.	Подання роботи на перевірку на плагіат	05.05.2025	
11.	Захист курсової роботи		

Студентка Третяк Є.М.

Керівник Борозенний С.О.

«_____» _____

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NLP - Natural Language Processing (обробка природної мови)

NLTK - Natural Language Toolkit

SVM - Support Vector Machine

АНОТАЦІЯ

У ході виконання даної роботи було проведено ознайомлення та дослідження основних методів машинного навчання та використання оптимального методу для досягнення найкращих результатів, розглянуто уже існуючі NLP-моделі для семантичного аналізу тексту, а також проведена робота із обробкою та використанням даних.

Під час дослідження було розглянуто вже наявні рекомендаційні системи та їхні підходи до надання користувачу рекомендацій для визначення нових підходів до створення системи.

Як результат роботи, було створено вебзастосунок рекомендаційної системи фільмів, який за текстовим поясненням користувача надає список рекомендацій. Даний вебзастосунок створений за допомогою сучасного фреймворку Flask [1], а для опрацювання даних використовувалася бібліотека pandas [2].

ВСТУП

Актуальність теми. У сучасному світі кіноіндустрія є досить прибутковою та затребуваною нішею. Було створено багато стримінгових сервісів, які користуються попитом завдяки своїй зручній функціональності та запитом користувача.

Зараз користувач не обмежений у своїх вподобаннях та можливостях перегляду кінопродукту, оскільки із розвитком інформаційних технологій було створено велику кількість сервісів, які націлені на допомогу користувачу переглянути будь-який продукт на основі його побажань, а кіноіндустрія розвивається щороку, створюючи величезну кількість нових фільмів.

Проте маючи таку велику різноманітність продуктів, користувачу може бути досить складно обрати саме ті фільми, які відповідають його вподобанням чи поточному настрою. Саме тому для покращення досвіду користувача різноманітні стримінгові сервіси почали впроваджувати рекомендаційні системи.

Уже наявні рекомендаційні системи здебільшого базуються на історії переглядів користувача та пошуку інших глядачів, які мають схожі вподобання. Зокрема, такий метод уже запровадили найбільші стримінгові компанії, такі як Netflix . Проте готових рішень стосовно розробки системи, яка б аналізувала настрій користувача та на основі цього надавала рекомендації, не було виявлено на етапі створення даного вебзастосунку.

Вибір теми дослідження було обґрунтовано потребою користувачів у новому представленні рекомендаційної системи, котра б слугувала додатковим чинником при виборі наступного фільму для перегляду у тому випадку, коли користувач не задоволений результатами уже існуючих систем. Рекомендаційна система зможе поліпшити досвід глядача та таким чином приносити прибутки для стримінгових сервісів та кінокомпаній.

Мета та завдання роботи. Метою даної роботи є дослідження наявних сервісів та ідея створення рекомендаційної системи фільмів на основі психологічного стану користувача для визначення топ-фільмів, які можуть сподобатися глядачу у даний момент.

Для досягнення поставленої мети були виконано наступні кроки:

- Проаналізовано основні методи, які наразі використовуються у найпопулярніших рекомендаційних системах.
- Запропонований підхід до створення системи, яка надає рекомандації на основі настрою глядача.
- Створено вебзастосунок рекомендаційної системи.
- як жанри впливають на емоції

Об’єкт дослідження. Процес пошуку та аналізу даних з використанням бібліотеки для їх обробки.

Предмет дослідження. Основні вибрані методи машинного навчання для класифікаці, уже наявні моделі для класифікації тексту та основні методи NLP, такі як лематизація, усунення слів без значення тощо.

Структура та обсяг курсової роботи. Робота складається зі вступу, чотирьох розділів, загальних висновків даної роботи, списку використаної літератури. Загальний обсяг курсової роботи становить 37 сторінок.

РОЗДІЛ 1
ОГЛЯД НАЯВНИХ РІШЕНЬ

Для кращого розуміння інтересів користувачів у даній предметній області, завжди хорошою ідеєю є аналіз уже готових рішень. Даний підхід дозволяє спробувати себе в ролі користувача та зрозуміти основні недоліки таких програм.

Було знайдено та проаналізовано наступні застосунки:

1. MovieLens
2. Letterboxd

1.1 MovieLens

MovieLens [3] - одна із найвідоміших рекомендаційних систем фільмів *(див. рисунок 1.2)*.

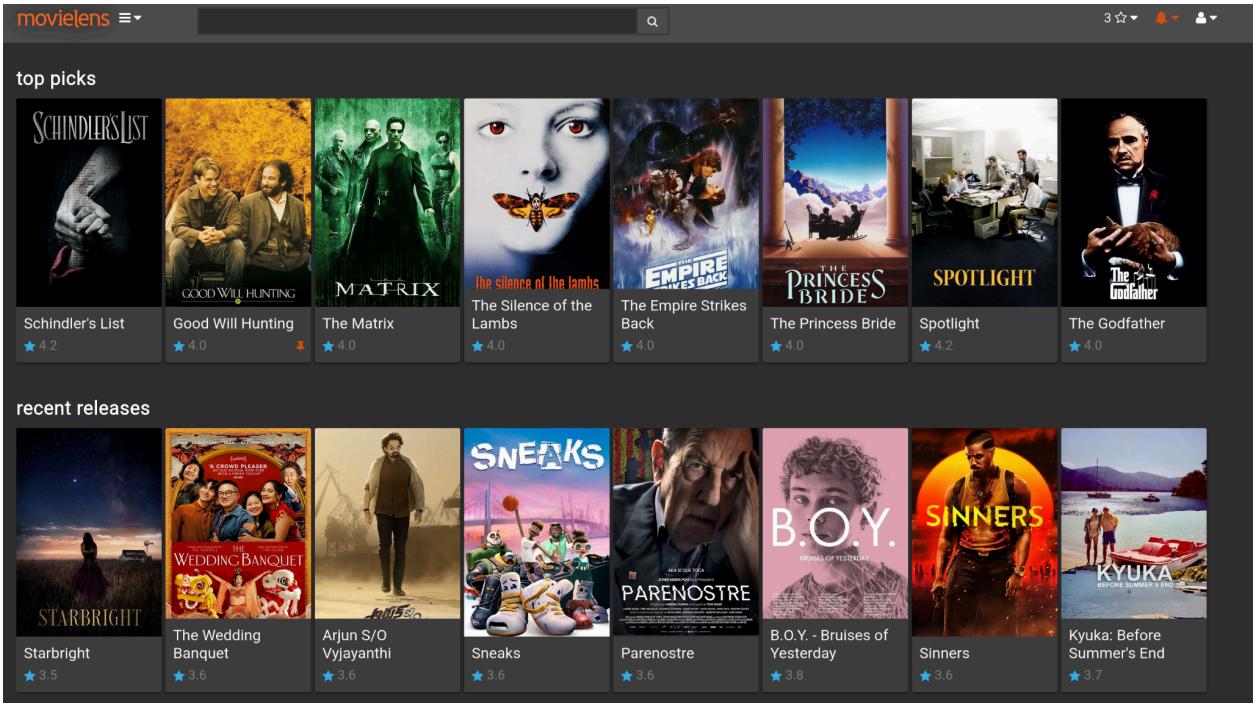

Рис. 1.2. Головна сторінка MovieLens [3]

Особливістю цього сервісу є те, що для отримання нових релевантних рекомендацій користувач має оцінювати вже переглянуті ним фільми. Таким чином, рекомендації можуть змінювати з часом, коли кількість відмічених користувачем фільмів буде збільшуватися.

Крім цього, даний сервіс містить доволі велику кінобібліотеку, що не обмежує глядача у його вподобаннях.

Даний застосунок має наступні можливості:

- Оцінювати вже переглянуті фільми (за шкалою від 1 до 5).
- Здійснювати швидкий пошук фільмів.
- Сортувати фільми за їх жанром, датою випуску, оцінкою.
- Додавати фільми до списку бажаних фільмів.
- Перегляд основної інформації про фільм.

Загалом, функціонал даного сервісу є доволі інтуїтивним, що робить досвід користувача приємнішим, а також не містить непотрібні функції, які б перезавантажували користувачів зайвою інформацією.

1.2 Letterboxd

Letterboxd [4] - це застосунок, який скоріше нагадує соціальну мережу для глядачів (*див. рисунок 1.2)*. На даному сервісі рекомендації надаються на основі вподобань інших користувачів, між якими є спільні інтереси.

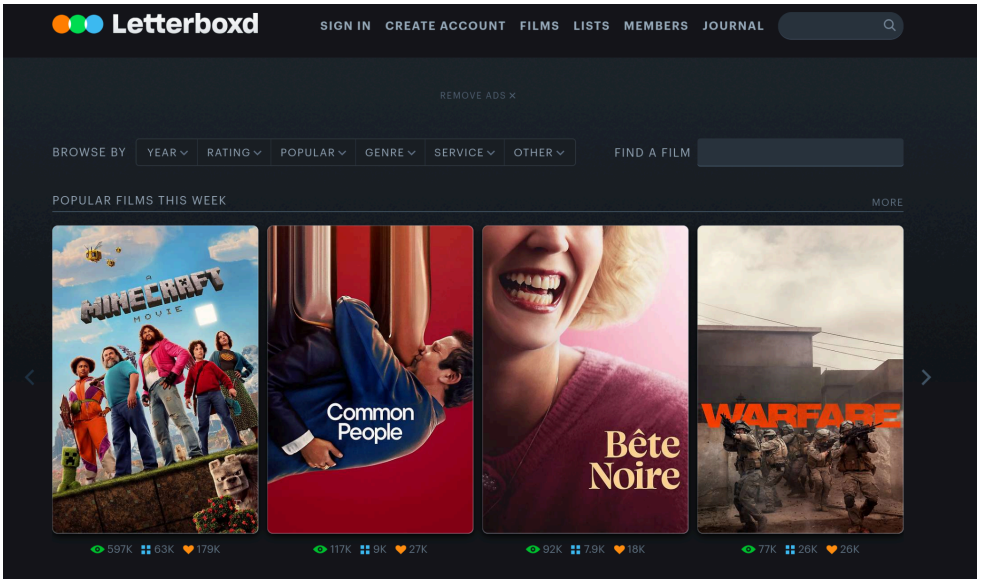

Рисунок 1.2. Головна сторінка Letterboxd [4]

Даний застосунок містить багато можливостей, а саме:

- Переглядати підбірки (тут їх дуже багато, реалізовані різноманітні списки фільмів за багатьма параметрами).
- Слідкувати за переглянутими фільмами інших глядачів.
- Сортувати фільми за різними параметрами.
- Реалізований пошук фільмів за назвою.

До недоліків або ж ідей для поліпшення даної розробки можна віднести те, що тут немає персоналізованої підбірки, яка б опиралася суто на вподобання користувача та його історію переглядів.

РОЗДІЛ 2
ЗБІР ТА ПІДГОТОВКА ДАНИХ

2.1. Пошук та аналіз датасету з фільмами

Для рекомендаційної системи важливим етапом розробки є саме дані. У цьому випадку необхідно було знайти датасет з фільмами, який би містив усю необхідну інформацію для створення рекомендацій та був би достатньо великого розміру для різноманіття. Саме з цього основного датасету будуть надаватися варіанти фільмів для користувача.

Під час пошуку підходящого датасету були занотовані та виконані наступні вимоги до нього:

- Мусить містити описи фільмів, оскільки ця інформація буде пізніше використана для визначення емоційного відтінку кожного із фільмів.
- Мусить містити назви, рейтинги, рік випуску фільмів, оскільки ця інформація є ключовою для користувача.
- Мусить містити інформацію про жанри фільмів. Така характеристика фільмів буде пізніше використана для аналізу емоційного відтінку фільмів.
- Має бути не дуже малим за розміром (бажано не менше, ніж декілька десятків тисяч екземплярів).

Для пошуку датасетів був використаний ресурс Kaggle [5]. Kaggle - це сервіс для пошуку багатьох датасетів, які пізніше використовуються для аналізу даних, практики з обробки великих даних та тренування моделей. Датасети там дуже різноманітні, наприклад, вони різних розмірів та різних сфер застосування.

Після аналізу найбільш популярних датасетів з фільмами, було обрано «The Movies Dataset» [5]. Ці дані відповідають вище переліченим вимогам, а саме:

- Містить більше 45000 екземплярів, що є достатнім розміром для даного студентського проєкту.
- Містить усю необхідну інформацію, а саме: перелік жанрів кожного із фільмів, назву мовою оригіналу, короткий опис, країни виробництва, тривалість, загальний рейтинг від глядачів тощо.

Єдиним недоліком цього датасету є те, що його дані не оновлюються. І тут присутні лише ті твори, які були випущені до липня 2017 року включно. Це може негативно вплинути на досвід користування рекомендаційної системи, яка буде заснована на цих даних, оскільки частка користувачів, можливо, буде мати бажання переглядати лише ті фільми, що були зняті впродовж останніх років.

2.2. Обробка головного датасету з фільмами

Перед використанням датасету необхідно завжди проводити його попередній аналіз та препроцесинг. Ці кроки необхідні для того, аби позбутися якихось аномалій у даних, які можуть пізніше негативно вплинути на якість даних або ж спотворити результати, а також привести датасет до універсального вигляду, з яким потім буде зручно працювати.

Чистка та препроцесинг датасету у стандартному вигляді включає наступні кроки:

- Визначити, які дані відсутні і прийняти рішення стосовно їхньої відсутності. Якщо у певній колонці дуже велика кількість відсутніх даних (орієнтовно більше 70 відсотків від кількості рядків), то одним із рішень є видалення цієї колонки. Проте якщо

дані дуже важливі для подальшої роботи, є також варіант заповнити цю колонку якимись значеннями.

- Привести тип кожної колонки до відповідного. Цей крок важливий, якщо ми далі хочемо працювати зі значеннями в колонках.
- Визначити стратегію стосовно дублікатів. Оскільки дуже часто дані є можуть містити помилки та повторні значення, у практиці надають перевагу видаленню дублікатів (наявність дублікатів може визначатися як від лише декількох колонок, так і від всього рядка, залежно від предметної області та поставленого завдання).
- Знайти викиди у даних (outliers). Це такі дані, які не можуть бути істинними у даній ситуації. Тому при наявності таких даних приймається рішення стосовно їх подальшої обробки.

Варто зазначити, що залежно від обраного датасету, предметної області та цільової мети, можуть проводитися додаткові кроки в процесі очищення даних та приведення їх до готового вигляду.

Після огляду датасету «The Movies Dataset» були виконані такі дії з очищення:

- Видалено колонку з інформацією про колекцію, до якої належить даний кінотвір. Ця ідея обґрунтована тим, що дана колонка містить близько 90% порожніх значень, і не несе жодної цінності при розробці.
- Видалено колонку з посиланнями на вебсторінки фільмів, оскільки тут було близько 82% порожніх значень, які неможливо заповнити якимись наперед визначенними значеннями чи згенерувати таке.
- Видалено колонку із значенням бюджету, витраченому на кінострічку. По перше, близько 80% рядків містять значення 0,

що не може відповідати дійсності у даній предметній області. По-друге, дана колонка не є важливою при розробці даного проєкту.

- Видалення дублікатів.
- Видалення тих рядків, де колонка із описом фільму є порожньою. Дана дія виправдана тим, що для складання емоційного профілю фільму дана колонка відіграє дуже велику роль.

Нижче наведено статистику щодо відсутніх значень у колонках перед очищенням *(див. рисунок 1.3)*.

	column_name	percent_missing
adult	adult	0.000000
belongs_to_collection	belongs_to_collection	90.115691
budget	budget	0.000000
genres	genres	0.000000
homepage	homepage	82.883913
id	id	0.000000
imdb_id	imdb_id	0.037391
original_language	original_language	0.024194
original_title	original_title	0.000000
overview	overview	2.098271
popularity	popularity	0.010997
poster_path	poster_path	0.848986
production_companies	production_companies	0.006598
production_countries	production_countries	0.006598
release_date	release_date	0.191352
revenue	revenue	0.013197
runtime	runtime	0.578454
spoken_languages	spoken_languages	0.013197
status	status	0.191352
tagline	tagline	55.104914
title	title	0.013197
video	video	0.013197
vote_average	vote_average	0.013197
vote_count	vote_count	0.013197

Рисунок 1.3. Статистика порожніх значень в датасеті «The Movies Dataset»

На рисунку можна побачити вище описані характеристики, а саме: велика кількість відсутньої інформації щодо колекції фільмів (колонка belongs_to_collection) та вебсторінки (колонка homepage).

2.3. Пошук та обробка допоміжних датасетів

Для визначення емоційного профілю кожного із фільмів було вирішено натренувати модель машинного навчання (що буде описано детальніше у наступних розділах). Проте після пошуків вже наявного датасету із фільмами та інформацією про їхні емоції, було виявлено, що немає підходящого датасету, який би задовільняв ці потреби. Під час пошуків було знайдено лише один такий датасет, проте через його специфіку даних було критично мало (лише близько 2 тисяч рядків).

Тому було вирішено знайти датасети лише із короткими описами фільмів, та створити свій власний датасет із відміткою про їхній емоційний відтінок (детальніше про цей процес буде описано пізніше).

Головний критерій при пошуку даних із виключно описами був саме розмір. Для якісних та більш точних результатів, звісно, бажано мати якнайбільше даних, відсутність чого є великою проблемою у сфері Data Science.

Вимоги до даних були задовільнені двома наступними датасетами:

- IMDb Movies/Shows with Descriptions (містить близько 8 тисяч екземплярів).
- Netflix Movies and TV Shows (містить орієнтовно 8 тисяч фільмів).

Для даного проєкту із цих двох датасетів була взята лише інформація із описами фільмів, хоча вони й містили також іншу інформацію.

Для обох датасетів (IMDb Movies/Shows with Descriptions та Netflix Movies and TV Shows) були виконані наступні дії для підготовки даних до пізнішого використання:

- Видалити дублікати (але орієнтуючись на дублікати саме в колонці із описами, оскільки це єдина потрібна колонка для подальшого роботи).
- Залишити лише колонку із описами.
- Видалити порожні значення та встановити правильні типи.

2.4. Опис використаної класифікаційної моделі

Наступним пунктом у роботі було визначити емоції фільмів для використання цих даних пізніше при тренуванні моделі.

Не зайвим буде надати коротке пояснення технологій, які будуть застосовуватися на даному кроці.

Одним із них є обробка природної мови (NLP) [6] - це сфера, котра використовує методи машинного навчання та інші підходи для розуміння людської мови. Зокрема, це використовується при розробці чат-ботів, аналізу текстів та автоматизації процесів, що вимагають розуміння тексту, написаного людиною.

Для завдання із визначення емоційного профілю буде проводитися семантичний аналіз тексту. Семантичний аналіз [7] - це завдання з розділу обробки природної мови (NLP) для визначення емоційного відтінку тексту. Такий аналіз може використовуватися, наприклад, для аналізу відгуків користувачів, коли необхідно зрозуміти загальне враження від певного продукту та автоматизувати процес визначення емоцій, які переважають у відгуках.

У даному випадку завдання полягає в тому, аби визначити переважаючу емоцію із опису фільму. Після проведення назначення таких міток, новий створений датасет буде використаний для тренування моделі.

Для проведення семантичного аналізу використовувалася дотренована transformer модель. Transformer - це архітектура глибинної нейронної мережі, котра вперше була запропонована компанією Google у 2017 році. Дана модель є однієї із найпотужніших на сьогодні й нерідко використовується для подальшого дотреновування на специфічних даних для конкретного завдання. Transformer моделі наразі знайшли своє використання у таких задачах, як: переклад тексту у режимі реального часу, проведення семантичного аналізу тощо [8].

Під час розробки була взята модель Fine-tuned DistilRoBERTa-base for Emotion Classification [10].

Основою цієї моделі слугувала модель Emotion English DistilRoBERTa-base, яка часто використовується як основа для створення нових, більш точних та специфічних моделей, які б давали точніші результати для конкретного завдання з класифікації [11].

Модель Emotion English DistilRoBERTa-base [11] використовується для визначення переважаючих емоцій з тексту англійською мовою. Після аналізу тексту модель надає результат у шести класах емоцій (гнів, сум, відраза, страх, радість, здивування), а також нейтральний клас (котрий описує текст, який здебільшого не має сильно вираженого емоційного відтінку). Модель дає достатньо точні результати на текстах, які описують саме почуття та емоції людини.

Наприклад, якщо задати моделі короткий текст “I feel motivated today!” (перекладається як “Я почуваюся вмотивованим сьогодні!”), то вона покаже наступні результати, де переважаючою емоцією є саме радість (98% із загальних 100%), що повністю відповідає дійсності (див. рисунок 2.1.).

```
from transformers import pipeline
classifier = pipeline("text-classification", model="j-hartmann/emotion-english-distilroberta-base", return_all_scores=True)
```

Show hidden output

```
[4] classifier("I feel motivated today!")
```

```
[[{'label': 'anger', 'score': 0.004516750108450651},
  {'label': 'disgust', 'score': 0.0015632084105163813},
  {'label': 'fear', 'score': 0.0007199379615485668},
  {'label': 'joy', 'score': 0.9886520504951477},
  {'label': 'neutral', 'score': 0.0021105345804244 28},
  {'label': 'sadness', 'score': 0.0013248475734144 45},
  {'label': 'surprise', 'score': 0.0011127226753160357}]]
```

Рисунок 2.1. Результати роботи моделі Emotion English DistilRoBERTa-base [11]

Для розробки рекомендаційної системи була використання поліпшена версія моделі Emotion English DistilRoBERTa-base, а саме: Fine-tuned DistilRoBERTa-base for Emotion Classification [10]. Поліпшення даної моделі полягає у тому, що вона була дотренована на нових даних, а саме на діалогах із популярних (переважно комедійних) телевізійних шоу, що дало змогу надати більше конкретики для кожного із класів емоцій. Даний підхід дав змогу поліпшити точність моделі та краще розрізняти запити із емоціями людей. Ця модель так само має шість класів емоцій та нейтральний клас.

Наприклад, при заданні промпту “I like people I met today” (перекладається як “Мені сподобалися люди, яких я зустрів/ла сьогодні!”) модель визначає, що переважаюча емоція у даному тексті радість (із точністю 99 відсотків) (див. рисунок 2.2) .

```
classifier = pipeline("sentiment-analysis", model="michellejieli/emotion_text_classifier")
classifier("I like people who I met today!")
```

```
[{'label': 'joy', 'score': 0.9904347658157349}]
```

Рисунок 2.2. Результати роботи моделі Fine-tuned DistilRoBERTa-base for Emotion Classification [10]

2.5. Використання класифікаційної моделі для доповнення датасету

Вище описана модель буде використовуватися для проставлення міток з переважаючою емоцією фільму на основі його короткого опису. Для цього було взято об’єднаний датасет, який містить лише описи фільмів та до нього була додана інформація про основну емоцію цих описів.

Проте при проставленні таких міток було вирішено уникати два класи та не записувати до результату нейтральний клас та клас здивування. Даний крок був оґрунтований тим, що коли модель аналізує опис фільму, то дуже часто переважаючою емоцією є саме нейтральний клас, оскільки в описі зазвичай згадуються події фільму, без чіткого емоційного забарвлення. Навіть людині нерідко важко зрозуміти, чи буде фільм сумний або радісний лише на основі його опису. До того ж, пізніше в розробці ми не зможемо опиратися на нейтральний клас, оскільки потрібно мати чіткі емоції, котрі допомогли б сформувати профіль фільму. Як результат, робота буде проводитися лише із п’ятьма емоціями (сум, радість, відраза, страх та гнів).

Тому при отриманні результатів від моделі обиралася та емоція, яка має найбільший відсоток попадань та не є нейтральною чи класом здивування.

Після роботи моделі із визначенням головної емоції було отримано доповнений датасет із лише двома колонками: описом та емоцією (див. рисунок 2.3).

description	emotion
As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.	sadness
After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth.	disgust
To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.	anger
Feuds, flirtations and toilet talk go down among the incarcerated women at the Orleans Justice Center in New Orleans on this gritty reality series.	joy
In a city of coaching centers known to train India’s finest collegiate minds, an earnest but unexceptional student and his friends navigate campus life.	joy
The arrival of a charismatic young priest brings glorious miracles, ominous mysteries and renewed religious fervor to a dying town desperate to believe.	joy
Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasi and Unicorns should be pals — and, hoof to heart, she’s determined to prove it.	joy
On a photo shoot in Ghana, an American model slips back in time, becomes enslaved on a plantation and bears witness to the agony of her ancestral past.	disgust
A talented batch of amateur bakers face off in a 10-week competition, whipping up their best dishes in the hopes of being named the U.K.'s best.	joy
A woman adjusting to life after a loss contends with a feisty bird that's taken over her garden — and a husband who's struggling to find a way forward.	disgust
Sicily boasts a bold "Anti-Mafia" coalition. But what happens when those trying to bring down organized crime are accused of being criminals themselves?	disgust

Рисунок 2.3. Результат доповненого датасету із використанням класифікаційної моделі

Даний датасет пізніше буде використаний для тренування моделі.

Також було проведено короткий аналіз на визначення збалансованості датасету (див. рисунок 2.4).

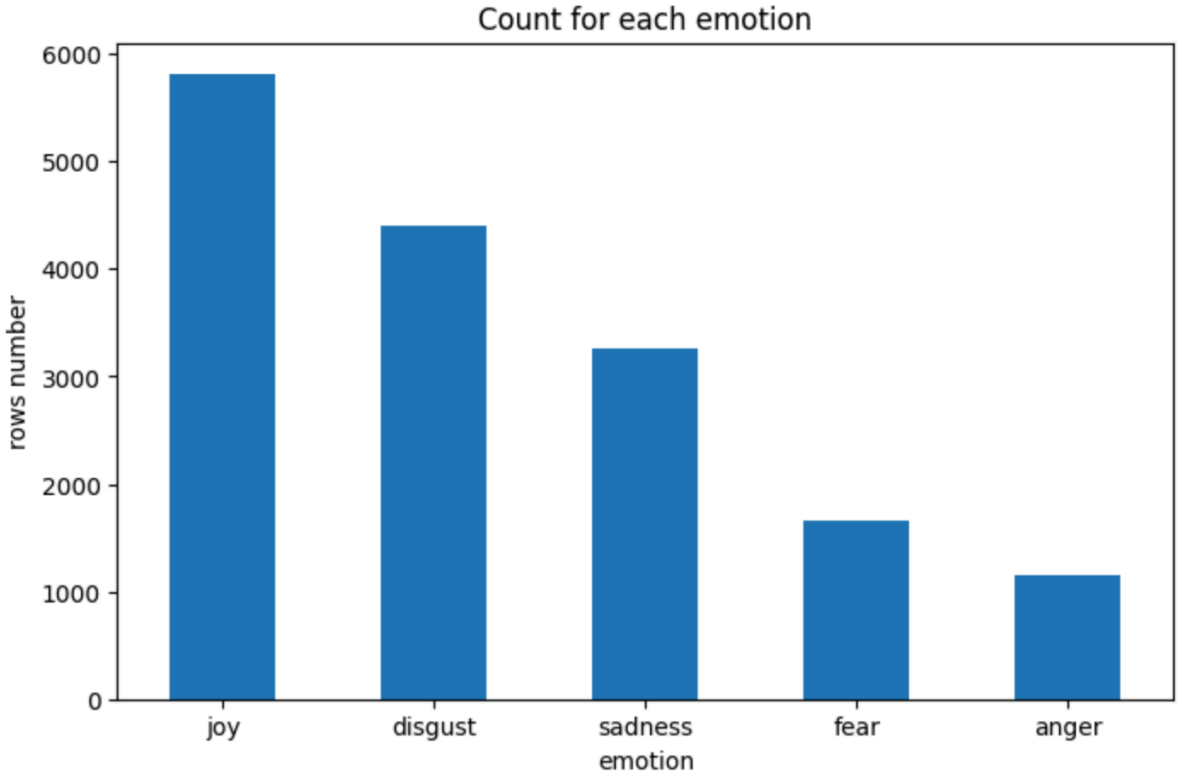

Рисунок 2.4. Кількість фільмів з кожною емоцією в доповненому датасеті

На графіку можна спостерігати, що найбільше фільмів є з позначкою “радість” та “відраза”, а найменше - із позначкою “гнів”, і датасет є незбалансованим.

2.6. Підготовка датасету з описами та емоціями фільмів до тренування

Датасет, який містить опис фільму та його переважаючу емоцію, буде використаний для тренування моделі, яка пізніше буде передбачати емоції для інших описів.

Перед тренуванням необхідно привести дані у цьому датасеті до потрібного формату, оскільки класичні методи машинного навчання не працюють з даними у звичайному текстовому вигляді.

Для препроцесингу текстових даних була використана бібліотека NLTK [12]. Дана бібліотека використовується для проведення семантичного аналізу та різних процесів обробки природної мови. Під час підготовки даних до тренування були виконані наступні дії:

- Приведення усіх символів до нижнього регістру.
- Видалення усіх символів, що не є літерами англійського алфавіту.
- Видалення стоп-слів. Стоп-слова в англійській мові - це, наприклад, артиклі a, the, an, а також сполучники (of, on, with, in), займенники тощо. Тобто це слова, які не несуть логічного значення в тексті, проте присутні через граматичні правила цієї мови. Для тренування такі дані створюють зайвий шум, тому правильним підходом буде позбутися їх.
- Проведення процесу лематизації для кожного слова. Лематизація - це процес приведення слова до його базової форми (іншими словами, знаходження леми), яка завжди є існуючим словом (тобто знаходиться у словнику). Наприклад, в українській мові:
    - Для слів “йшли”, “йде”, “йтиме” , лемою буде “іти”.
    - Для слова “книжок”, лемою буде “книжка” і т.д.

Лематизація у даній ситуації є стандартним пунктом на етапі препроцесингу тексту, що допомагає знизити кількість різноманітних слів та зосередитися на їхніх базових формах.

Після препроцесингу текстів було отримано наступні дані (див. рисунок 2.5.)

	description
2	jodie foster star clarice starling top student fbi training academy jack crawford scott glenn want clarice interview dr hannibal lecter a…
3	sequel set eleven year terminator young john connor edward furlong key civilization victory future robot uprising target shape shifting r…
4	disney animated feature follows adventure young lion simba jonathan taylor thomas heir father mufasa james earl jones simba wicked uncle …
5	vincent vega john travolta jules winnfield samuel l jackson hitman penchant philosophical discussion ultra hip multi strand crime movie s…
6	andy dufresne tim robbins sentenced two consecutive life term prison murder wife lover sentenced tough prison however andy know commit cr…
7	james cameron titanic epic action packed romance set ill fated maiden voyage r titanic pride joy white star line time largest moving obje…
8	victor johnny depp victoria emily watson family arranged marriage though like victor nervous ceremony forest practicing line wedding tree…
9	commodus joaquin phoenix take power strip rank maximus russell crowe one favored general predecessor father emperor marcus aurelius great…
10	robotic boy first programmed love david haley joel osment adopted test case cybertronics employee sam robards wife france connor though g…
11	examines one enduring civilization history
12	new decepticons appear optimus prime summons bumblebee give task saving earth help veteran autobot assembles rogue team young bot capture…
13	set mid nd century year james kirk helmed famous vessel installment star trek franchise set enterprise nx first earth starship capable wa…
14	murderous ventriloquist dummy terrorizes newlywed
15	life two canadian sister gail travers macha grenon changed mutual first love japanese pianist eiji okuda

Рисунок 2.5. Приклад обробленого тексту в процесі підготовки до тренування

РОЗДІЛ 3
НАВЧАННЯ ТА ЗАСТОСУВАННЯ МОДЕЛІ

3.1. Постановка задачі тренування моделі

Для передбачень емоцій фільму за його описом було вирішено використати попередньо зібрані та доповнені дані для тренування моделі машинного навчання. Дана модель потім буде використана для доповнення головного датасету з фільмами, де додана інформація полягатиме у визначенні переважаючих емоцій кожного із фільмів.

Оскільки завдання полягає у призначенні мітки класу емоції, то потрібно визначити найбільш оптимальний алгоритм машинного навчання для мультикласової класифікації із наявними даними, який буде давати найкращі результати. Мультикласова класифікація - це призначення вхідним даним лише один клас із декількох можливих (до того ж, класів більше, ніж два).

Для визначення алгоритму, котрий дає найкращі результати, було вирішено дослідити обрані алгоритми та обрати один із наступних: Logistic Regression, SVM.

3.2. Тренування моделі за допомогою алгоритму Logistic Regression

Алгоритм Logistic Regression доволі часто використовується для багатокласової класифікації [13]. На відміну від використання цього алгоритму для бінарної класифікації (коли дані передбачають лише два можливих вихідних класи), коли використовується сигмоїдна функція, для класифікації з багатьма класами використовується функція softmax. Завдання цієї функції полягає у тому, аби призначити кожному класу для поточних вхідних даних ймовірність того, що саме цей клас є правильним

передбаченням. Сума ймовірностей з усіх класів дорівнює одиниці, і функція обирає як правильне передбачення той клас, який має найбільшу ймовірність.

Даними, які модель має навчитися класифікувати - це описи фільмів, а вихідними класами є п’ять емоцій (гнів, відраза, страх, радість, сум).

Оскільки датасет є незбалансованим, під час тренування моделі застосовується параметр class_weight=’balanced’. Даний параметр необхідний для того, аби збільшити ваги помилок для класів, яких менше у тренувальних даних. Також для тренування виділяється 70 відсотків усіх даних, а на оцінку моделі 30 відсотків.

Після тренування моделі було визначено її показники точності, визначені на тестових даних (див. рисунок 3.1):

	precision	recall	f1-score	support
anger	0.33	0.32	0.33	351
disgust	0.55	0.57	0.56	1345
fear	0.46	0.45	0.45	488
joy	0.70	0.65	0.68	1771
sadness	0.49	0.52	0.50	928
accuracy			0.56	4883
macro avg	0.50	0.50	0.50	4883
weighted avg	0.57	0.56	0.56	4883

Accuracy: 0.5625639975424944

Рисунок 3.1. Показники точності моделі з використанням алгоритму Logistic Regression

З аналізу точності моделі видно, що загальна точність складає 56 відсотків, при цьому найгірше модель розпізнає клас гніву (що є очікуваним результатом, оскільки екземплярів цього класу найменше в даних). Також порівняно низька точність пояснюється тим, що визначити спектр емоцій з текстового опису фільму є доволі складним завдання, адже в описі зазвичай

можуть бути перелічені суто події або ж переказ фільму, без жодних виражених емоцій. Тому в даному випадку навіть людині це завдання з класифікацією емоцій буде нелегким, і, як наслідок, дана невисока точність допускається.

3.3. Тренування моделі за допомогою алгоритму SVM (Support Vector Machine)

Алгоритм SVM дуже часто застосовується для класифікаційних задач. Вважається, що зазвичай він дає кращі результати, ніж Logistic Regression, особливо у випадку із текстовими даними, а також менш сприятливий до перенавчання [14].

Під час тренування було вирішено змінювати гіперпараметри, аби знайти оптимальний варіант, який би давав найкращі результати.

1. Під час першої спроби було обрано наступні параметри: функцію втрат ‘hinge’ (це традиційна функція втрат для SVM); l2 регуляризацію (це додатковий штраф для того, аби модель не перенавчалась, контролює, чи не занадто великі ваги має модель); параметр alpha=10^-3 , чим більше це число, тим сильніша буде регуляризація; максимальна кількість ітерацій дорівнює п’яти (отже, буде не більше п’яти епох під час тренування); також додано балансування вагів для класів. (Див. рис. 3.2)

```
sgd = Pipeline([('vect', CountVectorizer()),
                ('tfidf', TfidfTransformer()),
                ('clf', SGDClassifier(loss='hinge', penalty='l2', alpha=1e-3,
                                      random_state=42, max_iter=5, tol=None,
                                      class_weight='balanced')),
               ])
```

Рисунок 3.2. Параметри для тренування першої SVM моделі

Після тренування було отримано наступні показники точності моделі (див. рис. 3.3).

```
accuracy 0.5525291828793775
              precision    recall  f1-score   support

       anger       0.27      0.42      0.33       347
     disgust       0.61      0.43      0.50      1329
        fear       0.40      0.61      0.48       473
         joy       0.69      0.69      0.69      1770
     sadness       0.52      0.50      0.51       964

    accuracy                           0.55      4883
   macro avg       0.50      0.53      0.50      4883
weighted avg       0.58      0.55      0.56      4883
```

Рисунок 3.3. Результати тренування першої SVM моделі

Загальна точність моделі складає 55 відсотків.

2. Під час другої спроби були змінені наступні параметри: функція втрат була обрана squared_hinge (схожа на функцію hinge, проте штраф зростає квадратично, а не лінійно, що змушує модель робити більш точніші передбачення); максимальна кількість епох дорівнює десяти. (Див. рис. 3.4)

```
sgd = Pipeline([('vect', CountVectorizer()),
                ('tfidf', TfidfTransformer()),
                ('clf', SGDClassifier(loss='squared_hinge', penalty='l2',
                                      alpha=1e-3, random_state=42, max_iter=10,
                                      tol=None, class_weight='balanced')),
               ])
```

Рисунок 3.4. Параметри для тренування другої SVM моделі

Дана модель була оцінена в 58 відсотків загальної точності (див. рис. 3.5.)

```
accuracy 0.58488634036453
              precision    recall  f1-score   support

       anger       0.36      0.35      0.35       347
     disgust       0.59      0.52      0.55      1329
        fear       0.47      0.49      0.48       473
         joy       0.68      0.73      0.70      1770
     sadness       0.53      0.54      0.54       964

    accuracy                           0.58      4883
   macro avg       0.53      0.53      0.53      4883
weighted avg       0.58      0.58      0.58      4883
```

Рисунок 3.5. Результати тренування другої SVM моделі

Після вищевказаних результатів було проведено ще декілька спроб покращити модель, змінивши її параметри, проте останній описаний варіант тренування виявився найкращим. У проєкті було вирішено використати алгоритм SVM із застосуванням найоптимальніших параметрів.

3.4. Застосування навченої моделі

Через те, що модель має невисокі показники із класифікацією класів емоцій фільму, було вирішено трохи змінити підхід до використання моделі. Замість отримання лише одного класу із переважаючою емоцією, у проєкті отримуються ймовірності призначення фільму кожного із п’яти класів емоцій. За допомогою такого підходу, можна визначити, які емоційні відтінки загалом присутні в описі фільму й скласти чіткішу картину.

Для складання емоційного профілю була також застосована інформація зі списку жанрів фільму й використана статистична інформація щодо вмісту емоцій у кожному із жанрів. (Див. рис. 3.6). Ці дані були проаналізовані за фільмами 2000-2021 років [15].

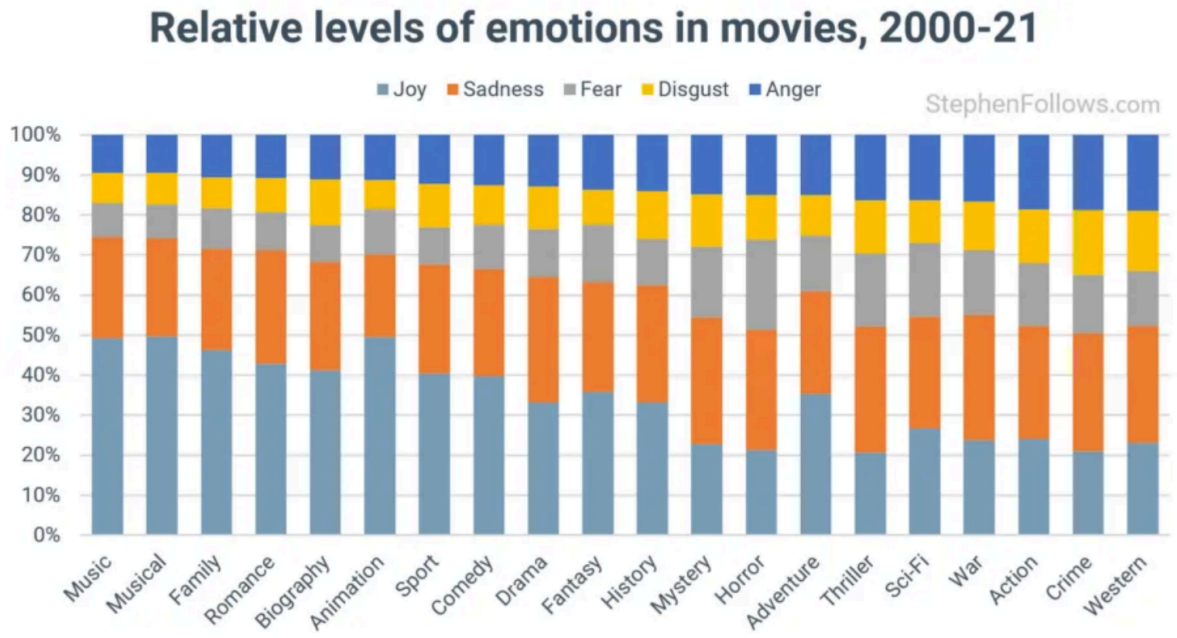

Рисунок 3.6. Статистичні дані щодо наявності кожної із п'яти емоцій у кожному жанрі [15]

Із цією інформацією було складено емоційний профіль фільмів лише за їхніми вказаними жанрами (у вигляді вектору із числовими значеннями, що відображають відсоткову наявність кожної із емоцій), а потім скомбіновано цей вектор із вектором, що відображає емоції, проаналізовані з коротких описів фільмів. Таким чином, був отриманий кінцевий варіант вектору, що містить загальну оцінку емоцій фільму. Саме цей фінальний вектор буде використовуватися для генерування рекомендацій користувачу.

РОЗДІЛ 4
ГЕНЕРУВАННЯ РЕКОМЕНДАЦІЙ

4.1. Обраний варіант генерування рекомендацій

Оскільки головною ідеєю генерування рекомендацій є знаходження фільмів, що не відповідають настрою користувача (наприклад, система має зважати на те, що якщо користувач сумний, то йому мають підбиратися оптимістичніші фільми), було застосовано наступний підхід: фінальний вектор із оцінкою емоцій фільму змінити на протилежний, тобто усі емоції будуть мати протилежні значення; після цього застосувати алгоритм для пошуку найбільш схожих векторів між настроєм користувача та фільмами.

Для пошуку фільмів із максимально схожими векторами було проаналізовано та використано алгоритм NearestNeighbors (пошук найближчих сусідів) [16]. Цей алгоритм дозволяє знайти зазначену кількість найбільш схожих векторів зі збірки даних. Аби удосконалити різноманіття рекомендацій, програма шукає певну кількість найближчих сусідів, проте користувачу надається лише певна частина цих рекомендацій, обрана випадковим чином. Такий підхід дозволяє урізноманітнити досвід користувача.

Також в процесі виконання проєкту було помічено, що незайвим було б надати додаткову інформацію щодо кожної із рекомендацій для користувача, яка відсутня у зібраних даних (наприклад, постер фільму та загальну оцінку, яка є релевантною на момент запиту). Тому при наданні рекомендацій користувачу надсилається запит до стороннього API, яке заповнює цю прогалину [17].

4.2. Приклад використання вебзастосунку

Для прикладу використання рекомендаційного вебзастосунку було надіслано такий опис почуттів користувача: “Я почуваюся дуже сумним останнім часом”. (Див. рис. 4.1)

Рисунок 4.1. Приклад запиту для розробленої рекомендаційної системи

Після відправлення цього запиту та отримання рекомендацій, користувач зможе переглянути рекомендації та коротку інформацію про фільм, такі як назва, посилання на ресурс для детальнішого ознайомлення, рік, загальний рейтинг, країни виробництва, жанри цієї кінострічки тощо (див. рис. 4.2 та рис. 4.3).

Оскільки користувач пояснив, що почувається сумним, частина рекомендованих кінострічок буде мати жанри комедій та більш розважальних фільмів для підняття настрою.

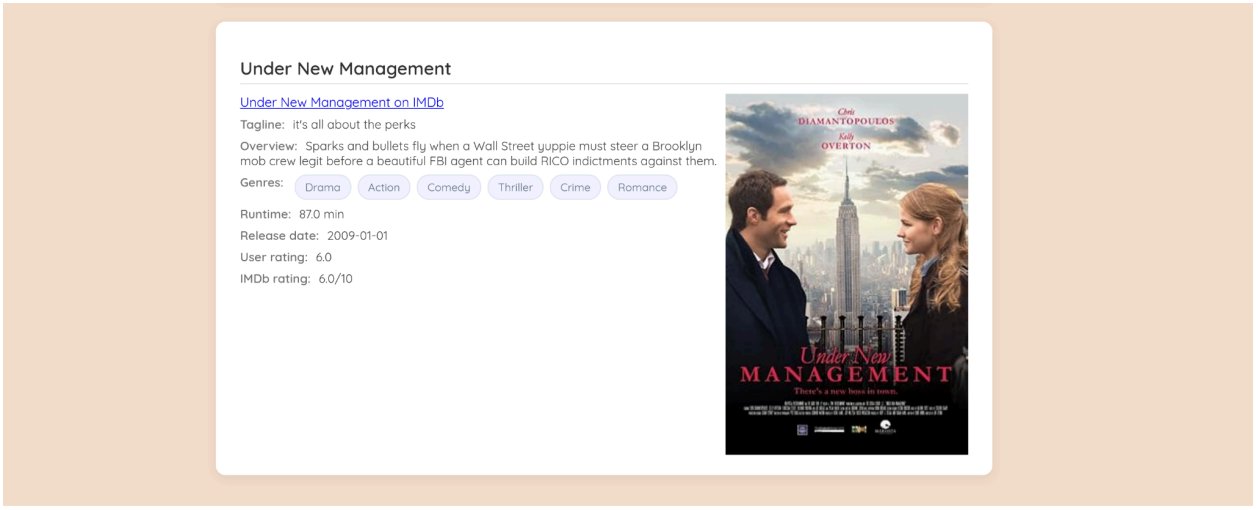

Рисунок 4.2. Приклад однієї з отриманих рекомендацій (I)

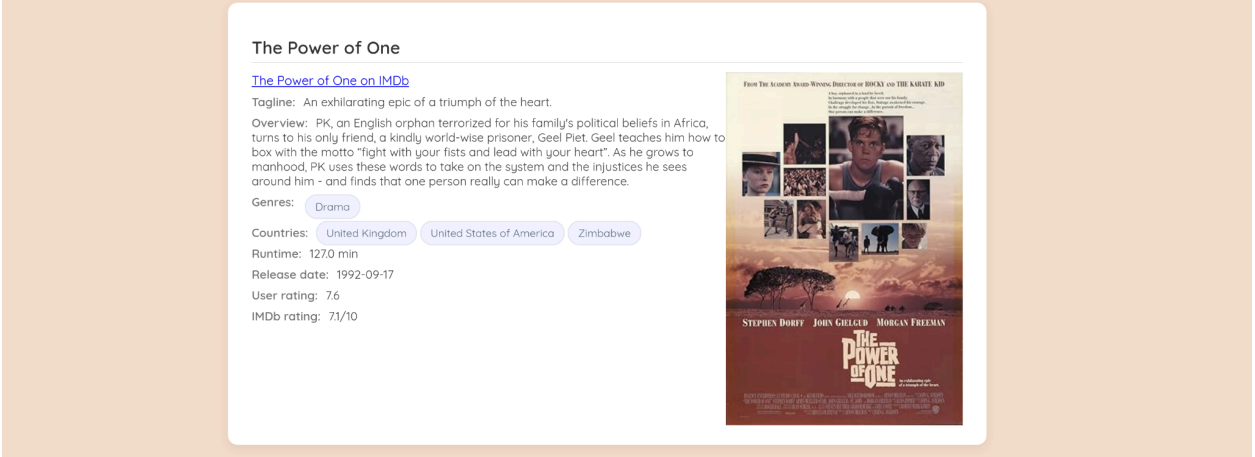

Рисунок 4.3. Приклад однієї з отриманих рекомендацій (II)

ВИСНОВКИ

Першим завданням під час виконання даної курсової роботи було провести пошук, аналіз, обробку та приведення попередньо відібраних даних до потрібного формату для подальшого їх використання.

Було проаналізовано вже наявні рекомендаційні системи фільмів, визначено їх недоліки та переваги для загального розуміння потреб користувача.

Мету завдання досягнуто, оскільки створено вебзастосунок рекомендаційної системи фільмів, що за інформацією про емоційний стан користувача надає список рекомендованих кінострічок. Даний підхід таким чином зможе поліпшити психологічне самопочуття користувача та підвищити перегляди кінострічок, що позитивно впливатимуть на прибутки кіноіндустрії.

Під час розробки було проведено тренування та використання моделі машинного навчання, проведення семантичного аналізу емоцій кінострічок та психологічного профілю користувача (тобто його переважаючих емоцій на даний момент). Також було обрано та використано алгоритм NearestNeighbors для обрання фільмів на етапі генеруванні списку рекомендацій.

Під час обробки даних були застосовані методи із розділу NLP для підготовки даних до тренування та приведення їх до коректного формату.

Після закінчення розробки застосунок був протестований на відповідність до загальних умов виконання проєкту.

Проте, даний вебзастосунок також містить деякі недоліки, які можуть бути усунуті у подальших версіях проєкту. До таких недоліків, зокрема, належать: обмежена кількість кінострічок, які пропонуються користувачу, а також певна неактуальність цих даних (оскільки дані містять фільми лише до 2017 року включно); відсутність додаткових вказівок від користувача стосовно надання йому рекомендацій (наприклад, користувач не має

можливості вказати, що бажає зараз кінострічки лише певного жанру чи певних країн виробництва). Даний проєкт може слугувати відправною точкою для подальших удосконалень.

Під час розробки були набуті практичні навички з використання фреймворку Flask [1] для створення вебінтерфейсу користувача з дотриманням основних норм налаштування проєкту за допомогою цього фреймворку. Крім цього, було створено власний дизайн вебзастосунку, що гарантує позитивний досвід користувача від використання даної програми.

СПИСОК ЛІТЕРАТУРИ

1. Flask [Електронний ресурс] – Режим доступу до ресурсу: https://flask.palletsprojects.com/en/stable/
2. Pandas [Електронний ресурс] – Режим доступу до ресурсу: https://pandas.pydata.org/
3. MovieLens [Електронний ресурс] – Режим доступу до ресурсу: https://movielens.org/
4. Letterboxd [Електронний ресурс] – Режим доступу до ресурсу: https://letterboxd.com/
5. Kaggle. The Movies Dataset [Електронний ресурс] / Kaggle – Режим доступу до ресурсу: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
6. IBM. What is NLP (natural language processing)? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/think/topics/natural-language-processing
7. IBM. What is sentiment analysis? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/think/topics/sentiment-analysis
8. Rick Merritt. What is A Transformer Model? [Електронний ресурс] / Rick Merritt – 2022 – Режим доступу до ресурсу: https://blogs.nvidia.com/blog/what-is-a-transformer-model/
9. IBM. What is a transformer model? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/think/topics/transformer-model
10. Li Michelle. Fine-tuned DistilRoBERTa-base for Emotion Classification [Електронний ресурс] / Hugging Face – Режим доступу до ресурсу: https://huggingface.co/michellejieli/emotion_text_classifier

11. Hartmann Jochen. Emotion English DistilRoBERTa-base [Електронний ресурс] / Hugging Face – 2022 – Режим доступу до ресурсу: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base
12. Introduction to NLTK: Tokenization, Stemming, Lemmatization, POS Tagging [Електронний ресурс] – Режим доступу до ресурсу: https://www.geeksforgeeks.org/introduction-to-nltk-tokenization-stemming-lemmatization-pos-tagging/
13. IBM. What is logistic regression? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/think/topics/logistic-regression
14. IBM. What are support vector machines (SVMs)? [Електронний ресурс] / IBM – Режим доступу до ресурсу: https://www.ibm.com/think/topics/support-vector-machine
15. Understanding movie genre emotions [Електронний ресурс] / Stephen Follows – Режим доступу до ресурсу: https://stephenfollows.com/p/understanding-movie-genres-emotions
16. Nearest Neighbors [Електронний ресурс] / scikit-learn – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html#sklearn.neighbors.NearestNeighbors
17. The Open Movie Database [Електронний ресурс] / OMDb API – Режим доступу до ресурсу: https://www.omdbapi.com/