



**Artificial Intelligence – Spring 2025**

# **Semester Project Filmception**

**An AI-powered Multilingual movie summary  
translator and genre classifier**

**Group members:** Sabreena Azhar, Bakhtawar Ilyas

**Roll number:** i221751, i221665

**Submission Date:** 5/7/2025



## **Table of Contents**

<input type="checkbox"/> <b>Introduction</b> .....	3
<input type="checkbox"/> <b>Steps</b> .....	3
Part 1: DATA COLLECTION AND LOADING .....	3
Part 2: TEXT TRANSLATION AND AUDIO CONVERSION:.....	4
Part 3: TEXT PREPROCESSING .....	6
Part 4: MOVIE GENRE PREDICTION MODEL.....	7
Part 5: CHAT INTERFACE .....	11
Part 6: Model Evaluation .....	17
Evaluation Metrics Used.....	17
Code Workflow.....	17
<input type="checkbox"/> <b>Summary</b> .....	24
<input type="checkbox"/> <b>References</b> .....	Error! Bookmark not defined.



## National University of Computer and Emerging Sciences Islamabad Campus

---

### • Introduction

This project allows users to input a movie summary and explore two main features: **genre prediction** and **multilingual audio narration**.

To build this system, we used the **CMU Movie Summary Dataset** and applied thorough **text preprocessing**, including cleaning, tokenization, and lemmatization. The genre prediction task is handled using a **Logistic Regression** model trained on **TF-IDF** features, tackling the challenge of multi-label classification.

To enhance accessibility, the system translates each movie summary into **Arabic, Urdu, and Korean** and then converts it into speech using a **Text-to-Speech engine**. This lets users listen to summaries in their preferred language.

With a user-friendly **menu-based interface**, FilmCeption offers:

- **Audio playback**
- **Genre prediction**

### • Steps

#### Part 1: DATA COLLECTION AND LOADING

##### Data Source:

Dataset: <https://www.kaggle.com/datasets/msafi04/movies-genre-dataset-cmu-movie>

We downloaded dataset from kaggle.

"movie.metadata.tsv"

"plot\_summaries.txt"

Loaded these files in first step.



# National University of Computer and Emerging Sciences

## Islamabad Campus

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** localhost:8888/lab/tree/Untitled%20Folder%20/maincode1.ipynb
- File Menu:** File, Edit, View, Run, Kernel, Tabs, Settings, Help
- Toolbar:** +, FilmCeptior X, maincode1.X, backend.py X, movie.meta X, character.m X, raw\_merge X, movie\_tran X
- Code Cell:** Contains Python code for predicting movie genres based on user input. It uses NLTK resources and processes movie metadata.
- Output Cell:** Shows the output of the code execution, including NLTK initialization and processing logs.
- Bottom Status Bar:** Mode: Command, Ln 31, Col 2, maincode1.ipynb, 1, 1:00 AM, 5/7/2025, 24°C, ENG

## Part 2: TEXT TRANSLATION AND AUDIO CONVERSION:

In this phase of the FilmCeption project, we focused on making movie summaries multilingual and accessible through speech. Specifically, we translated at least 50 movie summaries into Korean, Arabic, and Urdu, and then converted these translations into audio files using Text-to-Speech (TTS) technology.

Using the GoogleTranslator API, we implemented a robust translation function with retry logic to handle failures and limit input length. The translated texts were saved in a structured format inside raw\_translations.csv for future reference.

Next, we used gTTS (Google Text-to-Speech) to convert these translated summaries into audio. Each audio file was saved in the audio\_output directory, organized by language and movie ID. The function included retry logic and file size checks to ensure quality and reliability.

We wrapped everything into a pipeline (process\_translations) that handled:

- Translating summaries into three languages.
- Saving translated text files.



# National University of Computer and Emerging Sciences

## Islamabad Campus

- Generating and storing corresponding MP3 audio files.

```
print("\nActual Genres:")
for genre in movie_row['genres']:
    print(f"- {genre}")

elif choice == '4':
    print("Exit")
    break
else:
    print("Invalid choice")

if __name__ == "__main__":
    main()

Initializing NLTK resources...
FILECEPTION , BY SABREENA AND BAKHTAWAR
Loading raw data...
Loading metadata...
Loading plot summaries...
Processing genres...
Merging data...

Processing translations for movie summaries...
Translating:  0%| t/s] | 0/50 [00:00<?, ?i
```



# National University of Computer and Emerging Sciences

## Islamabad Campus

The screenshot shows a Jupyter Notebook interface running on a local server at `localhost:8888/lab/tree/Untitled%20Folder%202`. The left sidebar displays a file tree for an unnamed folder containing several files: `audio_output`, `data`, `character.metadata`, `maincode1.ipynb` (selected), `movie.metadata.tsv`, `name.clusters.txt`, `plot_summaries.txt`, `README.txt`, and `tvtropes.clusters.txt`. The right pane shows the code for `maincode1.ipynb` and its execution output. The code includes logic for handling user input, processing NLTK resources, loading raw data, and translating movie summaries. A progress bar indicates the completion of a translation task. The bottom status bar shows the kernel is Python 3 (ipykernel) and the current time is 11:31 PM on 5/6/2025.

### Part 3: TEXT PREPROCESSING

In this part of the FilmCeption project, we focused on cleaning and preparing the movie summary data for effective genre prediction.

We started by implementing a `clean_text()` function that handled:

- Lowercasing the text for standardization.
- Removing special characters, numbers, and punctuation using regular expressions.
- Tokenizing the text into individual words.
- Removing stopwords and short words that don't contribute meaningful information.
- Lemmatizing words to reduce them to their base forms for consistency.

This ensured the text was standardized, noise-free, and ready for machine learning.

Next, in the `preprocess_data()` function, we:

- Applied text cleaning to every movie summary.
- Extracted and filtered multi-label genres from the metadata.
- Removed rare genres that appeared fewer than 50 times to improve model focus.
- Filtered out any entries that ended up with no genres after the filtering.



# National University of Computer and Emerging Sciences

## Islamabad Campus

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ip X movie.metadata X character.metadata X raw\_merged\_d X movie\_translat X + Notebook Python 3 (ipykernel)

Name Modified

audio\_output 2m ago

data 2m ago

character.metadata 3y ago

maincode1.ipynb 1m ago

movie.metadata.tsv 3y ago

name.clusters.txt 3y ago

plot\_summaries.txt 3y ago

README.txt 3y ago

tvtropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

```
elif choice == '4':
    print("Exit")
    break
else:
    print("Invalid choice")

if __name__ == "__main__":
    main()

Initializing NLTK resources...
FILMCEPTION , BY SABREENA AND BAKHTAWAR
Loading raw data...
Loading metadata...
Loading plot summaries...
Processing genres...
Merging data...

Processing translations for movie summaries...
Translating: 100%|██████████| 50/50 [25:29<00:00, 30.58
s/it]

Preprocessing data...
Preprocessing text...
Cleaning: 74%|██████████| 30865/41793 [02:13<01:00, 181.50i
t/s]
```

Simple Python 3 (ipykernel) | Busy Mode: Edit Ln 5, Col 11 maincode1.ipynb 1 11:31 PM 5/6/2025

Type here to search

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ip X movie.metadata X character.metadata X raw\_merged\_d X movie\_translat X + Notebook Python 3 (ipykernel)

Name Modified

audio\_output 2m ago

data now

character.metadata 3y ago

maincode1.ipynb 1m ago

movie.metadata.tsv 3y ago

name.clusters.txt 3y ago

plot\_summaries.txt 3y ago

README.txt 3y ago

tvtropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

```
break
else:
    print("Invalid choice")

if __name__ == "__main__":
    main()

Initializing NLTK resources...
FILMCEPTION , BY SABREENA AND BAKHTAWAR
Loading raw data...
Loading metadata...
Loading plot summaries...
Processing genres...
Merging data...

Processing translations for movie summaries...
Translating: 100%|██████████| 50/50 [25:29<00:00, 30.58
s/it]

Preprocessing data...
Preprocessing text...
Cleaning: 100%|██████████| 41793/41793 [03:00<00:00, 231.88i
t/s]
Filtering genres...

Loaded 41761 movies with genres
Creating train/test split...
```

Simple Python 3 (ipykernel) | Busy Mode: Edit Ln 5, Col 11 maincode1.ipynb 1 11:32 PM 5/6/2025

Type here to search

## Part 4: MOVIE GENRE PREDICTION MODEL

In this phase of the **FilmCeption** project, we built a **multi-label classification model** to predict a movie's genres based on its cleaned summary.



# National University of Computer and Emerging Sciences

## Islamabad Campus

*What We Implemented:*

- **Feature Extraction:**

We used **TF-IDF vectorization** to convert cleaned summaries into numerical features. The TF-IDF model captured the importance of words across summaries, using both unigrams and bigrams (1–2 word phrases) while filtering stopwords.

- **Model Selection:**

We opted for a **Logistic Regression model** wrapped in a **One-vs-Rest classifier**, which is well-suited for multi-label problems. It trains one classifier per genre label, enabling the model to assign multiple genres to a single movie.

- **Training and Testing:**

The dataset was split into training and test sets. The model was trained on the TF-IDF features of the training data and evaluated using the test set.

- **Evaluation:**

We generated a detailed classification report, showing metrics like precision, recall, and F1-score for each genre to assess the model's performance.

- **Persistence:**

The trained model, TF-IDF vectorizer, and multi-label binarizer were saved using joblib for future predictions and reuse.

- **Genre Prediction Function:**

We created a predict\_genres() function that accepts a movie summary and returns predicted genres after preprocessing and vectorizing the input.

```
if __name__ == "__main__":
    main()

Initializing NLTK resources...
FILMCEPTION , BY SABREENA AND BAKHTAWAR
Loading raw data...
Loading metadata...
Loading plot summaries...
Processing genres...
Merging data...

Processing translations for movie summaries...
Translating: 100%|██████████| 50/50 [25:29<00:00, 30.58 s/it]

Preprocessing data...
Preprocessing text...
Cleaning: 100%|██████████| 41793/41793 [03:00<00:00, 231.88i t/s]
Filtering genres...

Loaded 41761 movies with genres
Creating train/test split...

== Training Genre Prediction Model ==
Extracting features with TF-IDF...
Training model...
```



# National University of Computer and Emerging Sciences

## Islamabad Campus

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ipynb movie.meta X character.meta X raw\_merged\_c X movie\_translat X Python 3 (ipykernel)

== Training Genre Prediction Model ==  
Extracting features with TF-IDF...  
Training model...

Model Evaluation:

	precision	recall	f1-score	support
Absurdism	0.00	0.00	0.00	13
Action	0.42	0.76	0.54	1281
Action Comedy	0.06	0.20	0.10	35
Action Thrillers	0.12	0.46	0.19	91
Action/Adventure	0.31	0.76	0.44	700
Adult	0.11	0.38	0.16	23
Adventure	0.30	0.70	0.42	663
Adventure Comedy	0.07	0.24	0.11	21
Airplanes and airports	0.11	0.62	0.18	8
Alien Film	0.19	0.64	0.29	22
Americana	0.02	0.87	0.03	14
Animal Picture	0.14	0.46	0.22	28
Animated cartoon	0.05	0.29	0.08	17
Animation	0.42	0.79	0.55	489
Anime	0.20	0.57	0.30	63
Art film	0.09	0.26	0.13	84
Auto racing	0.31	0.75	0.44	12
Avant-garde	0.11	0.11	0.11	19
B-movie	0.09	0.44	0.15	66
Biographical film	0.10	0.47	0.17	188
Biography	0.12	0.36	0.19	124

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

Simple 2 Python 3 (ipykernel) | Busy Mode: Command Ln 498, Col 43 maincode1.ipynb 1 11:33 PM 5/6/2025

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ipynb movie.meta X character.meta X raw\_merged\_c X movie\_translat X Python 3 (ipykernel)

Black-and-white 0.28 0.65 0.39 765  
Blaxploitation 0.17 0.38 0.24 16  
Bollywood 0.24 0.77 0.36 211  
Boxing 0.09 0.30 0.14 10  
Buddy film 0.07 0.29 0.11 68  
Caper story 0.15 0.44 0.23 18  
Chase Movie 0.02 0.85 0.03 22  
Childhood Drama 0.07 0.24 0.11 25  
Children's 0.12 0.54 0.28 68  
Children's Fantasy 0.09 0.48 0.15 33  
Children's/Family 0.15 0.55 0.23 139  
Chinese Movies 0.38 0.68 0.49 174  
Christian film 0.16 0.43 0.24 23  
Christmas movie 0.15 0.50 0.23 10  
Combat Films 0.17 0.61 0.26 18  
Comedy 0.50 0.71 0.59 2084  
Comedy Thriller 0.00 0.00 0.00 13  
Comedy Western 0.05 0.18 0.08 11  
Comedy film 0.11 0.41 0.18 376  
Comedy of Errors 0.04 0.10 0.05 41  
Comedy of manners 0.02 0.07 0.03 43  
Comedy-drama 0.10 0.44 0.16 266  
Coming of age 0.11 0.48 0.18 149  
Computer Animation 0.10 0.38 0.15 34  
Costume Adventure 0.07 0.16 0.10 19  
Costume Horror 0.18 0.50 0.27 12  
Costume drama 0.21 0.68 0.32 80  
Courtroom Drama 0.21 0.51 0.29 43

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

Simple 2 Python 3 (ipykernel) | Busy Mode: Command Ln 498, Col 43 maincode1.ipynb 1 11:34 PM 5/6/2025



# National University of Computer and Emerging Sciences

## Islamabad Campus

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ip X movie.metadata X character.met X raw\_merged\_d X movie\_translat X + Notebook Python 3 (ipykernel)

Name / Untitled Folder 2 / Modified

audio\_output 2m ago  
data 1s ago  
character.metadat... 3y ago  
maincode1.ipynb now  
movie.metadata.tsv 3y ago  
name.clusters.txt 3y ago  
plot\_summaries.txt 3y ago  
README.txt 3y ago  
tvropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

Genre	Score 1	Score 2	Score 3	Count
Docudrama	0.06	0.24	0.09	54
Documentary	0.39	0.77	0.52	257
Domestic Comedy	0.08	0.18	0.11	49
Doomsday film	0.12	0.55	0.20	20
Drama	0.67	0.72	0.69	3818
Dystopia	0.04	0.18	0.05	20
Ensemble film	0.06	0.22	0.18	76
Epic	0.16	0.62	0.25	37
Erotic Drama	0.10	0.26	0.14	23
Erotic thriller	0.06	0.28	0.18	29
Erotica	0.09	0.27	0.14	44
Experimental film	0.05	0.18	0.07	21
Family Drama	0.10	0.46	0.16	147
Family Film	0.32	0.71	0.44	626
Family-Oriented Adventure	0.09	0.36	0.14	33
Fantasy	0.22	0.60	0.33	379
Fantasy Adventure	0.12	0.59	0.19	34
Fantasy Comedy	0.01	0.07	0.02	28
Feminist Film	0.04	0.07	0.05	14
Filipino Movies	0.05	0.12	0.07	42
Film adaptation	0.09	0.41	0.14	238
Film noir	0.14	0.54	0.23	115
Gangster Film	0.18	0.51	0.27	78
Gay	0.14	0.35	0.20	52
Gay Interest	0.15	0.37	0.21	52
Gay Themed	0.20	0.44	0.27	63
Gothic Film	0.06	0.14	0.09	14
Gross-out	0.06	0.22	0.10	6

Simple 2 Python 3 (ipykernel) | Busy Mode: Command L 498, Col 43 maincode1.ipynb 1 11:35 PM 5/6/2025

localhost:8888/lab/tree/Untitled%20Folder%202

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCeption.ip X maincode1.ip X movie.metadata X character.met X raw\_merged\_d X movie\_translat X + Notebook Python 3 (ipykernel)

Name / Untitled Folder 2 / Modified

audio\_output 2m ago  
data 1s ago  
character.metadat... 3y ago  
maincode1.ipynb now  
movie.metadata.tsv 3y ago  
name.clusters.txt 3y ago  
plot\_summaries.txt 3y ago  
README.txt 3y ago  
tvropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

Genre	Score 1	Score 2	Score 3	Count
Jungle Film	0.38	0.75	0.50	16
LGBT	0.18	0.54	0.27	170
Marriage Drama	0.06	0.27	0.09	22
Martial Arts film	0.34	0.77	0.47	145
Media Satire	0.16	0.29	0.21	17
Medical fiction	0.10	0.56	0.17	9
Melodrama	0.04	0.21	0.07	189
Mockumentary	0.11	0.16	0.13	25
Monster	0.26	0.48	0.34	23
Monster movie	0.12	0.38	0.18	32
Music	0.12	0.41	0.19	73
Musical	0.20	0.59	0.29	474
Musical Drama	0.02	0.11	0.03	18
Musical comedy	0.06	0.18	0.09	34
Mystery	0.22	0.63	0.33	430
Natural horror films	0.20	0.52	0.29	31
Neo-noir	0.04	0.20	0.07	10
New Hollywood	0.00	0.00	0.00	15
Parody	0.09	0.34	0.14	166
Period piece	0.14	0.53	0.22	253
Political cinema	0.12	0.34	0.18	38
Political drama	0.12	0.44	0.18	98
Political satire	0.15	0.24	0.18	21
Political thriller	0.14	0.40	0.20	40
Pornographic movie	0.12	0.38	0.18	29
Pre-Code	0.04	0.10	0.05	30
Prison	0.19	0.58	0.29	19
Propaganda film	0.04	0.11	0.05	18

Simple 2 Python 3 (ipykernel) | Busy Mode: Command L 498, Col 43 maincode1.ipynb 1 11:35 PM 5/6/2025



# National University of Computer and Emerging Sciences

## Islamabad Campus

The screenshot shows a Jupyter Notebook interface with several tabs open. The main code cell displays a table of movie genre classification data:

Genre	Sports	0.37	0.72	0.48	145
Spy	0.26	0.89	0.49	62	
Stop motion	0.10	0.36	0.16	28	
Superhero	0.22	0.61	0.32	18	
Superhero movie	0.22	0.60	0.32	45	
Supernatural	0.17	0.62	0.27	119	
Surrealism	0.05	0.10	0.06	21	
Suspense	0.10	0.46	0.17	128	
Swashbuckler films	0.22	0.65	0.33	23	
Tamil cinema	0.02	0.10	0.03	10	
Teen	0.18	0.62	0.28	171	
Television movie	0.04	0.25	0.07	115	
Thriller	0.42	0.73	0.53	1317	
Time travel	0.03	0.09	0.04	11	
Tragedy	0.03	0.06	0.04	17	
Tragicomedy	0.00	0.00	0.00	12	
War film	0.41	0.78	0.54	302	
Western	0.50	0.81	0.62	215	
Workplace Comedy	0.06	0.12	0.08	16	
World cinema	0.28	0.64	0.39	969	
Wuxia	0.15	0.64	0.24	11	
Zombie Film	0.39	0.50	0.44	42	
micro avg	0.26	0.63	0.37	29784	
macro avg	0.15	0.41	0.21	29784	
weighted avg	0.33	0.63	0.42	29784	
samples avg	0.27	0.65	0.35	29784	

## Part 5: CHAT INTERFACE

Our FilmCeption Project has a **console-based interactive interface** that allows users to engage with the system's features intuitively.

*Main Functionalities:*

1. **Data Loading & Initialization**
  - o Loads either the **translated** or **raw movie data**.
  - o Processes summaries into cleaned formats.
  - o Loads or creates the **train/test split**.
  - o Trains or loads the **genre classification model**.

2. **Interactive Menu Options:**

1. Translate a movie summary
2. Play existing audio translations
3. Predict genres from user-inputted summaries
4. Exit

*Feature Highlights:*

- **Translation & Text-to-Speech:**



# National University of Computer and Emerging Sciences

## Islamabad Campus

- Users can choose any movie\_id, select a target language (e.g., Urdu, Arabic) and generate both **text** and **MP3 audio** of the translated summary.
- Automatically plays the audio if generation succeeds.
- **Audio Playback:**
  - Lists all available .mp3 translation files.
  - Lets users select and play them from the terminal.
- **Genre Prediction:**
  - Accepts a custom movie summary from the user.
  - Cleans and vectorizes the input, then predicts genres using the trained **Logistic Regression model**.
  - If the summary exists in the original dataset, it also displays the **actual genres** for comparison.

Selecting choice=1

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook has several tabs open, including 'requirements.txt', 'FilmCception.ipynb', 'maincode1.ipynb' (which is the active tab), 'movie.metadata', 'character.metadata', 'raw\_merged.csv', and 'movie\_translat'. The main code cell in 'maincode1.ipynb' displays a table of genre distribution statistics and a main menu:

	Swashbuckler films	0.22	0.65	0.33	23
Tamil cinema	0.02	0.10	0.03	10	
Teen	0.18	0.62	0.28	171	
Television movie	0.04	0.25	0.07	115	
Thriller	0.42	0.73	0.53	1317	
Time travel	0.03	0.09	0.04	11	
Tragedy	0.03	0.06	0.04	17	
Tragicomedy	0.00	0.00	0.00	12	
War film	0.41	0.78	0.54	382	
Western	0.50	0.81	0.62	215	
Workplace Comedy	0.06	0.12	0.08	16	
World cinema	0.28	0.64	0.39	969	
Wuxia	0.15	0.64	0.24	11	
Zombie Film	0.39	0.50	0.44	42	
micro avg	0.26	0.63	0.37	29784	
macro avg	0.15	0.41	0.21	29784	
weighted avg	0.33	0.63	0.42	29784	
samples avg	0.27	0.65	0.35	29784	

```
==== Main Menu ====
1. Translate a movie (any ID)
2. Play existing translations
3. Predict genres from summary
4. Exit
Enter choice (1-4):
```



# National University of Computer and Emerging Sciences

## Islamabad Campus

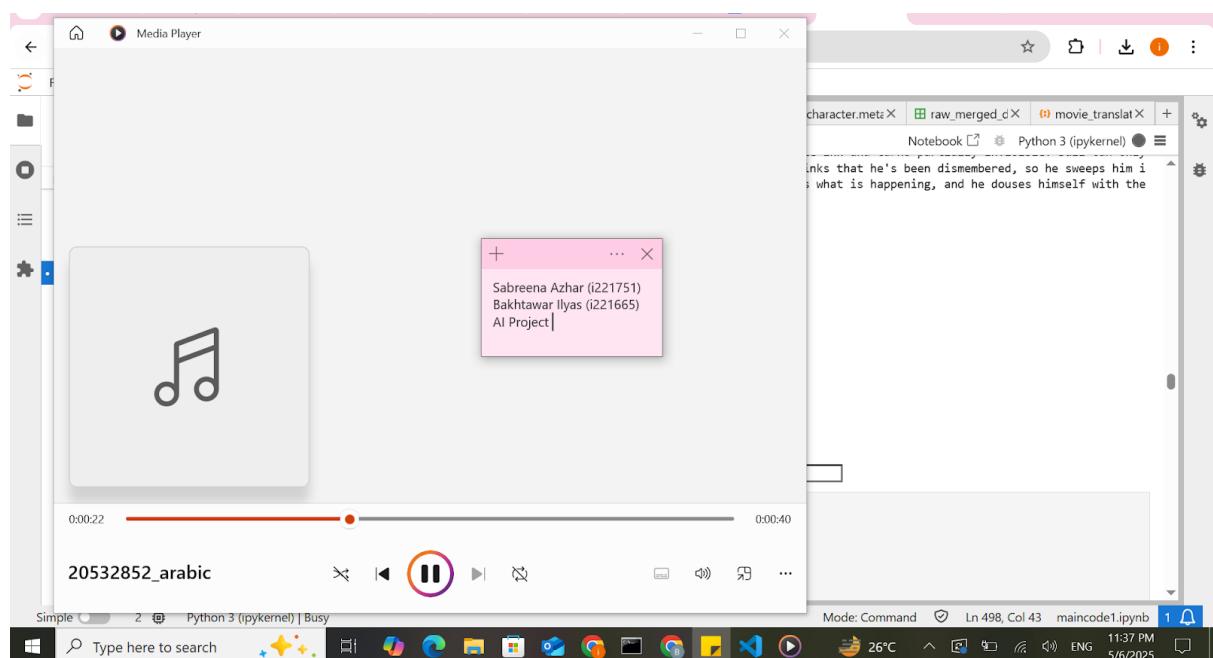
The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook is titled 'maincode1.ipynb'. In the code cell, the user has run a script that translates a movie summary from English to Arabic. The output text is:

```
... persuades him to merge with his threat to kill himself. At this point in the movie, a reference to Jesus Christ is made, that a historical "John Doe" has already died for the sake of humanity. The film ends with Connell turning to Norton and saying, "There you are, Norton! The people! Try and lick that!" 20532852
A line of people drool at the window of the shop of market butcher Buzz Buzzard. A short series of gags ensues about how Buzz dishonestly "jacks" up all his prices. Since Woody is broke as usual, he sneaks in and gets thrown out by Buzz. On the way out, Woody collides with a bottle of invisible ink and turns partially invisible. Buzz can only see parts of Woody's body, and, in a somewhat gruesome scene, thinks that into a trap door to get rid of him. When Woody awakes, he realizes what is rest of the ink in order to pose as a ghost.

Enter movie ID to translate: 20532852
Available languages:
ar: arabic
ur: urdu
ko: korean
Enter language code: ar
Saved arabic translation to audio_output\20532852_arabic.txt
Audio generated successfully

==== Main Menu ====
1. Translate a movie (any ID)
2. Play existing translations
3. Predict genres from summary
4. Exit
Enter choice (1-4):
```

The status bar at the bottom indicates 'Python 3 (ipykernel) | Busy'.



Selecting Choice =2



# National University of Computer and Emerging Sciences

## Islamabad Campus

```
localhost:8888/lab/tree/Untitled%20Folder%202
File Edit View Run Kernel Tabs Settings Help
requirements.t.txt FilmCeption.ipynb maincode1.ipynb movie.metadata character.metaz raw_merged_d movie_translat +
Notebook Python 3 (ipykernel)
+ - + ... ×
Sabreena Azhar (i221751)
Bakhtawar Ilyas (i221665)
AI Project |
```

==== Main Menu ====  
1. Translate a movie (any ID)  
2. Play existing translations  
3. Predict genres from summary  
4. Exit  
Enter choice (1-4): 2

Available audio files:  
1. 10799612\_arabic.mp3  
2. 10799612\_korean.mp3  
3. 10799612\_urdu.mp3  
4. 12978934\_arabic.mp3  
5. 12978934\_korean.mp3  
6. 12978934\_urdu.mp3  
7. 1335380\_arabic.mp3  
8. 1335380\_korean.mp3  
9. 1335380\_urdu.mp3  
10. 13455076\_arabic.mp3  
11. 13455076\_korean.mp3  
12. 13455076\_urdu.mp3  
13. 14330027\_arabic.mp3  
14. 14330027\_korean.mp3  
15. 14330027\_urdu.mp3  
16. 14880747\_arabic.mp3  
17. 14880747\_korean.mp3  
18. 14880747\_urdu.mp3  
19. 15072401\_arabic.mp3  
20. 15072401\_korean.mp3

Simple 2 Python 3 (ipykernel) Busy Mode: Command L 498, Col 43 maincode1.ipynb 1 11:38 PM 5/6/2025

Type here to search

```
localhost:8888/lab/tree/Untitled%20Folder%202
File Edit View Run Kernel Tabs Settings Help
requirements.t.txt FilmCeption.ipynb maincode1.ipynb movie.metadata character.metaz raw_merged_d movie_translat +
Notebook Python 3 (ipykernel)
+ - + ... ×
Sabreena Azhar (i221751)
Bakhtawar Ilyas (i221665)
AI Project |
```

132. 5272176\_urdu.mp3  
133. 5414895\_arabic.mp3  
134. 5414895\_korean.mp3  
135. 5414895\_urdu.mp3  
136. 5425205\_arabic.mp3  
137. 5425205\_korean.mp3  
138. 5425205\_urdu.mp3  
139. 595909\_arabic.mp3  
140. 595909\_korean.mp3  
141. 595909\_urdu.mp3  
142. 8388648\_arabic.mp3  
143. 8388648\_korean.mp3  
144. 8388648\_urdu.mp3  
145. 8471210\_arabic.mp3  
146. 8471210\_korean.mp3  
147. 8471210\_urdu.mp3  
148. 9252321\_arabic.mp3  
149. 9252321\_korean.mp3  
150. 9252321\_urdu.mp3

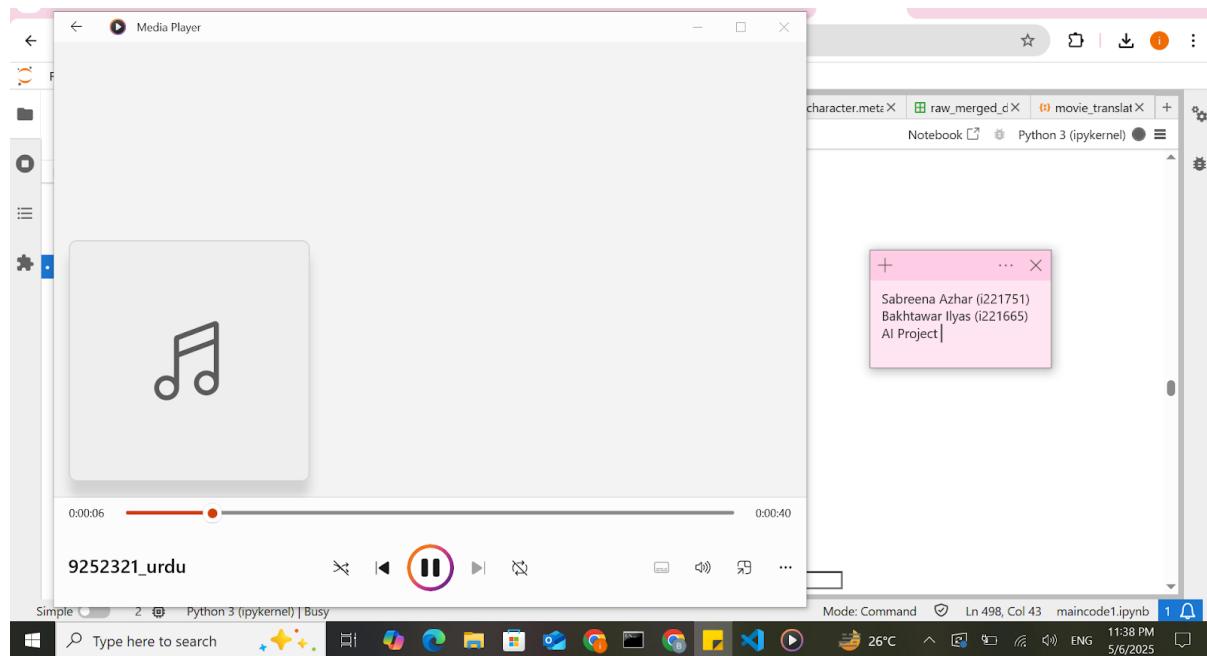
>Select file to play (or 'back'): 150

==== Main Menu ====  
1. Translate a movie (any ID)  
2. Play existing translations  
3. Predict genres from summary  
4. Exit  
Enter choice (1-4):

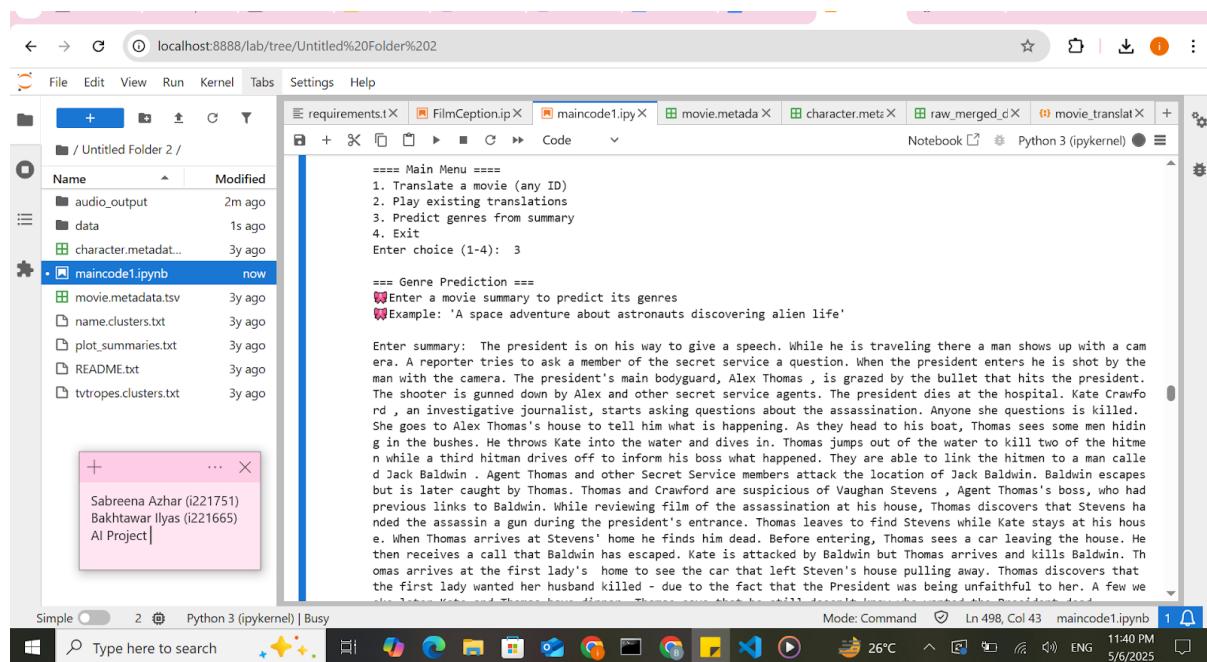
Simple 2 Python 3 (ipykernel) Busy Mode: Command L 498, Col 43 maincode1.ipynb 1 11:39 PM 5/6/2025

Type here to search

# National University of Computer and Emerging Sciences Islamabad Campus



Selecting Choice=3





# National University of Computer and Emerging Sciences

## Islamabad Campus

The screenshot shows a Jupyter Notebook interface with the URL `localhost:8888/lab/tree/Untitled%20Folder%202`. The left sidebar lists files in 'Untitled Folder 2': requirements.txt, FilmCeption.ipynb, maincode1.ipynb (selected), movie.metadata, character.metadata, raw\_merged.csv, and movie\_translat. The main code cell displays the following output:

```
Predicted Genres:  
- Action  
- Action/Adventure  
- Crime Fiction  
- Crime Thriller  
- Indie  
- Mystery  
- Political drama  
- Political thriller  
- Psychological thriller  
- Spy  
- Suspense  
- Thriller  
  
Actual Genres:  
- Thriller  
- Action/Adventure  
- Action  
- Drama  
  
==== Main Menu ====  
1. Translate a movie (any ID)  
2. Play existing translations  
3. Predict genres from summary  
4. Exit  
Enter choice (1-4):
```

The status bar at the bottom indicates 'Python 3 (ipykernel) | Busy'.

Selecting Choice=4 (Exit)

The screenshot shows the same Jupyter Notebook interface as the previous one. The user has typed '4' in the input field and pressed Enter. The output now includes the selected choice:

```
==== Main Menu ====  
1. Translate a movie (any ID)  
2. Play existing translations  
3. Predict genres from summary  
4. Exit  
Enter choice (1-4): 4  
Exit
```

The status bar at the bottom indicates 'Python 3 (ipykernel) | Idle'.



## National University of Computer and Emerging Sciences Islamabad Campus

### Part 6: MODEL EVALUATION

This section focuses on evaluating the performance of the **genre prediction model** trained earlier using industry-standard classification metrics for **multi-label text classification**.

#### Evaluation Metrics Used

Metric	Purpose
<b>Hamming Loss</b>	Measures the fraction of incorrect labels to the total number of labels
<b>Precision</b>	How many predicted genres were actually correct
<b>Recall</b>	How many actual genres were correctly predicted
<b>F1-Score</b>	Harmonic mean of Precision and Recall for balanced performance

---

#### Code Workflow

##### 1. Loading Components

- Loads test data (test\_data.csv) and the serialized:
  - Trained **Logistic Regression model**
  - **TF-IDF vectorizer**
  - **MultiLabelBinarizer** (for decoding multi-label predictions)

##### 2. Evaluating the Model

- Uses the trained model to predict genres for the test data.
- Transforms raw summaries into TF-IDF vectors and genres into binary labels.
- Calculates the following:
  - Hamming Loss
  - Precision (weighted)
  - Recall (weighted)
  - F1-Score (weighted)

##### 3. Interpreting Results

- If Hamming Loss < 0.5: Indicates solid multi-label classification performance.
- Precision, Recall, and F1-Score are presented with user-friendly interpretations.



# National University of Computer and Emerging Sciences

## Islamabad Campus

localhost:8888/lab/tree/Untitled%20Folder%20/maincode1.ipynb

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCception.ip X maincode1.ip X backend.py X movie.meta X character.m X raw\_merged X movie\_translat X +

Name Modified

- audio\_output 2m ago
- data 1s ago
- backend.py 8s ago
- character.metadat... 3y ago
- maincode1.ipynb now
- movie.metadata.tsv 3y ago
- name.clusters.txt 3y ago
- plot\_summaries.txt 3y ago
- README.txt 3y ago
- tvropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

```
def main():
    print("FilmCception Model Evaluation - Key Metrics 🎉")

    df_test, model, vectorizer, mnb = load_data_and_model()

    metrics = evaluate_model(df_test, model, vectorizer, mnb)

    if __name__ == "__main__":
        main()

🎉 FilmCception Model Evaluation - Key Metrics 🎉

== Key Evaluation Metrics ==
Hamming Loss: 0.0436
Weighted Precision: 0.3343
Weighted Recall: 0.6271
Weighted F1-Score: 0.4212

== Interpretation ==
🎉 Nice Work Bakhtawar and Sabreena : Hamming Loss < 0.5 indicates good multi-label performance

Precision of 0.33 means the model is 33% reliable when predicting a genre
Recall of 0.63 means the model finds 63% of relevant genres
F1-Score of 0.42 balances precision and recall performance
```

Simple Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 maincode1.ipynb 1 12:34 AM 5/7/2025

localhost:8888/lab/tree/Untitled%20Folder%20

File Edit View Run Kernel Tabs Settings Help

requirements.t X FilmCception.ip X maincode1.ip X movie.meta X character.m X raw\_merged\_c X movie\_translat X +

Name Modified

- audio\_output 2m ago
- data 1s ago
- character.metadat... 3y ago
- maincode1.ipynb now
- movie.metadata.tsv 3y ago
- name.clusters.txt 3y ago
- plot\_summaries.txt 3y ago
- README.txt 3y ago
- tvropes.clusters.txt 3y ago

Sabreena Azhar (i221751)  
Bakhtawar Ilyas (i221665)  
AI Project

	Sports	0.37	0.72	0.48	145
Spy	0.26	0.89	0.40	62	
Stop motion	0.10	0.36	0.16	28	
Superhero	0.22	0.61	0.32	18	
Superhero movie	0.22	0.60	0.32	45	
Supernatural	0.17	0.62	0.27	119	
Surrealism	0.05	0.10	0.06	21	
Suspense	0.10	0.46	0.17	128	
Swashbuckler films	0.22	0.65	0.33	23	
Tamil cinema	0.02	0.10	0.03	10	
Teen	0.18	0.62	0.28	171	
Television movie	0.04	0.25	0.07	115	
Thriller	0.42	0.73	0.53	1317	
Time travel	0.03	0.09	0.04	11	
Tragedy	0.03	0.06	0.04	17	
Tragicomedy	0.00	0.00	0.00	12	
War film	0.41	0.78	0.54	382	
Western	0.50	0.81	0.62	215	
Workplace Comedy	0.06	0.12	0.08	16	
World cinema	0.28	0.64	0.39	969	
Wuxia	0.15	0.64	0.24	11	
Zombie Film	0.39	0.50	0.44	42	
micro avg	0.26	0.63	0.37	29784	
macro avg	0.15	0.41	0.21	29784	
weighted avg	0.33	0.63	0.42	29784	
samples avg	0.27	0.65	0.35	29784	

Simple Python 3 (ipykernel) | Idle Mode: Command Ln 7, Col 21 maincode1.ipynb 1 11:43 PM 5/6/2025

**Confusion matrix:**



# National University of Computer and Emerging Sciences

## Islamabad Campus

localhost:8888/lab/tree/Untitled%20Folder%202/maincode1.ipynb

File Edit View Run Kernel Tabs Settings Help

FilmCptiX maincodi plot\_sum X backend.X movie.mx character X raw\_merc movie\_trz X +

Notebook Python 3 (ipykernel)

Name Modified

- character.metadata 3y ago
- cleaned\_movies.csv 2h ago
- genre\_model.pkl 2h ago
- mbl.pkl 2h ago
- movie.metadata.tsv 3y ago
- name.clusters.txt 3y ago
- plot\_summaries.txt 3y ago**
- raw\_translations.csv 2h ago
- README.txt 3y ago
- test\_data.csv 2h ago
- tfidf\_vectorizer.pkl 2h ago
- train\_data.csv 2h ago
- tvropes.clusters.txt 3y ago

FilmCpti Model Evaluation

== Evaluation Metrics ==

Hamming Loss: 0.0436  
Weighted Precision: 0.3343  
Weighted Recall: 0.6271  
Weighted F1-Score: 0.4212

== Classification Report ==

	precision	recall	f1-score	support
Absurdism	0.00	0.00	0.00	13
Action	0.42	0.76	0.54	1281
Action Comedy	0.06	0.20	0.10	35
Action Thrillers	0.12	0.46	0.19	91
Action/Adventure	0.31	0.76	0.44	700
Adult	0.11	0.30	0.16	23
Adventure	0.30	0.70	0.42	663
Adventure Comedy	0.07	0.24	0.11	21
Airplanes and airports	0.11	0.62	0.18	8
Alien Film	0.19	0.64	0.29	22
Americana	0.02	0.07	0.03	14
Animal Picture	0.14	0.46	0.22	28
Animated cartoon	0.05	0.29	0.08	17
Animation	0.42	0.79	0.55	489
Anime	0.20	0.57	0.30	63
Art film	0.09	0.26	0.13	84

Simple Python 3 (ipykernel) | Idle Mode: Command Ln 2, Col 23 maincode1.ipynb 1 3:08 AM 5/7/2025

Type here to search

localhost:8888/lab/tree/Untitled%20Folder%202/maincode1.ipynb

File Edit View Run Kernel Tabs Settings Help

FilmCptiX maincodi plot\_sum X backend.X movie.mx character X raw\_merc movie\_trz X +

Notebook Python 3 (ipykernel)

Name Modified

- character.metadata 3y ago
- cleaned\_movies.csv 2h ago
- genre\_model.pkl 2h ago
- mbl.pkl 2h ago
- movie.metadata.tsv 3y ago
- name.clusters.txt 3y ago
- plot\_summaries.txt 3y ago**
- raw\_translations.csv 2h ago
- README.txt 3y ago
- test\_data.csv 2h ago
- tfidf\_vectorizer.pkl 2h ago
- train\_data.csv 2h ago
- tvropes.clusters.txt 3y ago

samples avg 0.27 0.65 0.35 29704

Success! Hamming Loss is less than 0.5, which means your model performs well.

Confusion Matrix for Absurdism

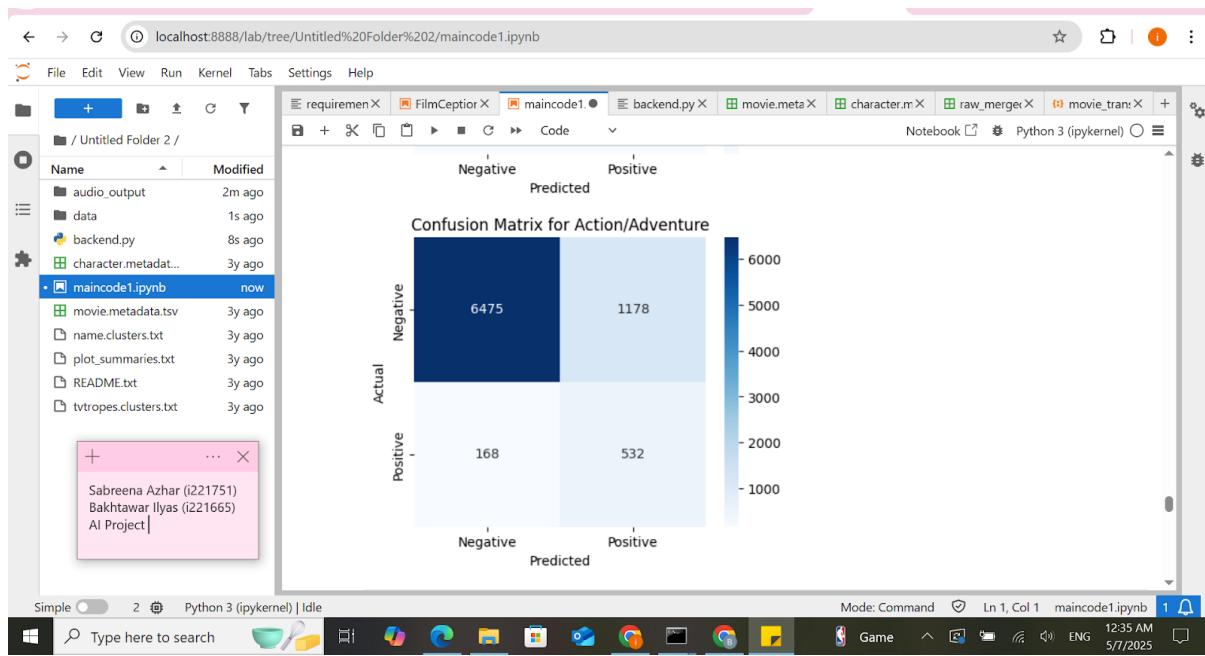
		Predicted	
		Negative	Positive
Actual	Negative	8319	21
	Positive	13	0

Mode: Command Ln 2, Col 23 maincode1.ipynb 1 3:09 AM 5/7/2025

Type here to search



# National University of Computer and Emerging Sciences Islamabad Campus



## Part 7: Interactive User Interface (GUI)

This section introduces the **Graphical User Interface of FilmCeption**, which transforms the project into an interactive, user-friendly application.

### Workflow Overview

Upon launching the GUI, users can navigate through these four main tabs:

#### 1. Translate Tab

*Translate a Movie Summary to Any Supported Language*

- User selects a **movie ID** and a **target language** (e.g., Arabic, Korean, Urdu).
- The system:
  - Translates the movie summary using Google Translate API.
  - Saves the translated text locally.
  - Offers an option to generate audio for it.



# National University of Computer and Emerging Sciences Islamabad Campus

The screenshot shows a Jupyter Notebook interface titled "FilmCaption - By Sabreena and Bakhtawar". The main window displays the "Filmception System , by Sabreena and Bakhtawar" tab. On the left, there is a file tree with various files and folders, including "plot\_summary" which is currently selected. The central part of the interface shows a table titled "Available Movies" with columns "ID" and "Summary". A dropdown menu indicates the "Target Language: Arabic (ar)". Below the table is a button labeled "Translate Selected". To the right, a "Translation Results" section shows a pink box containing the names of the two students: "Sabreena Azhar (i221751)" and "Bakhtawar Ilyas (i221665)" along with the text "AI Project". The bottom of the screen shows the Windows taskbar with various icons and the system tray indicating the date and time.

## 2. Audio Tools Tab

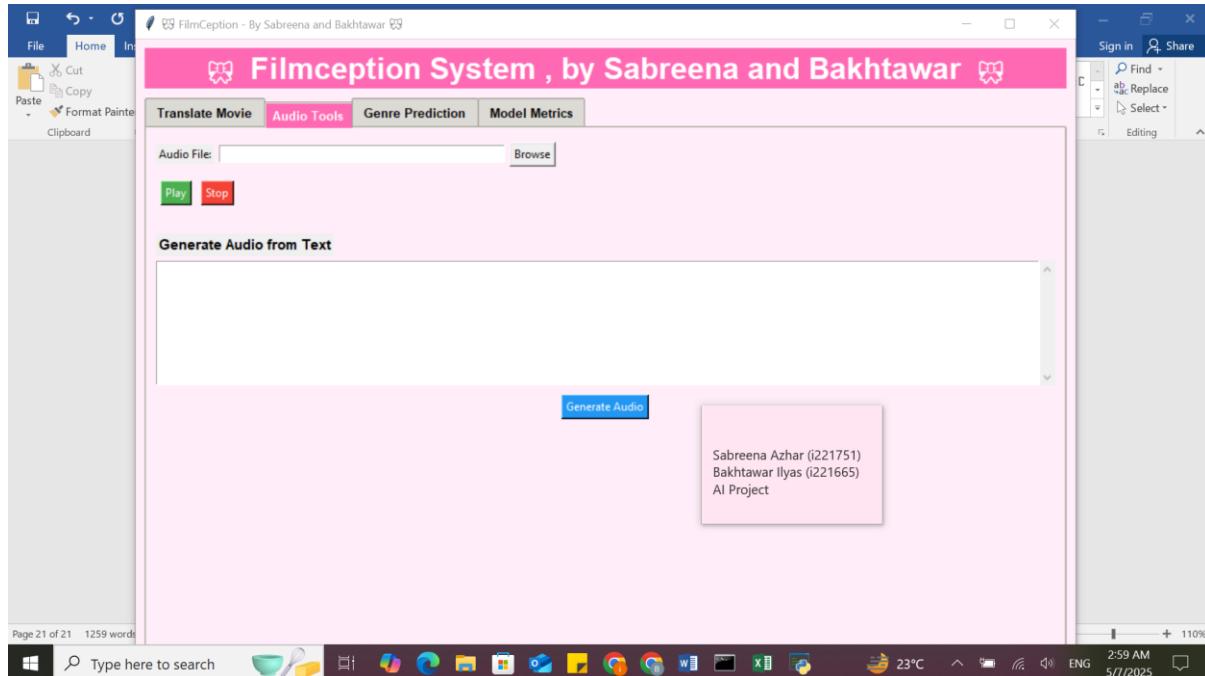
*Convert Summary to Speech and Listen*

- Users can:
  - **Play pre-generated audio files**
  - **Or generate new audio files** by selecting the movie and language
- Uses text-to-speech technology to create .mp3 files from summaries.



# National University of Computer and Emerging Sciences

## Islamabad Campus



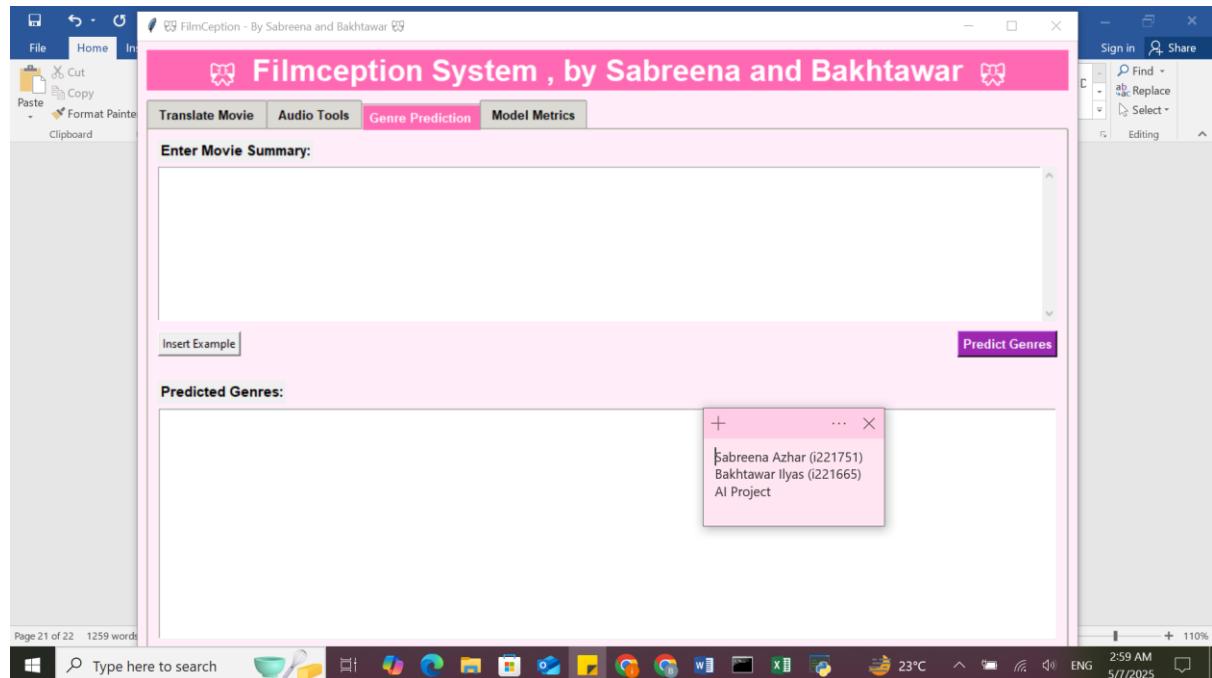
### 3. Genre Prediction Tab

Predict Movie Genre from User Input Summary

- User inputs any **movie summary** manually.
- The system uses the trained **Logistic Regression + TF-IDF model** to predict **multiple possible genres**.
- If the summary exists in the dataset, it also shows **actual genres** for comparison.



# National University of Computer and Emerging Sciences Islamabad Campus



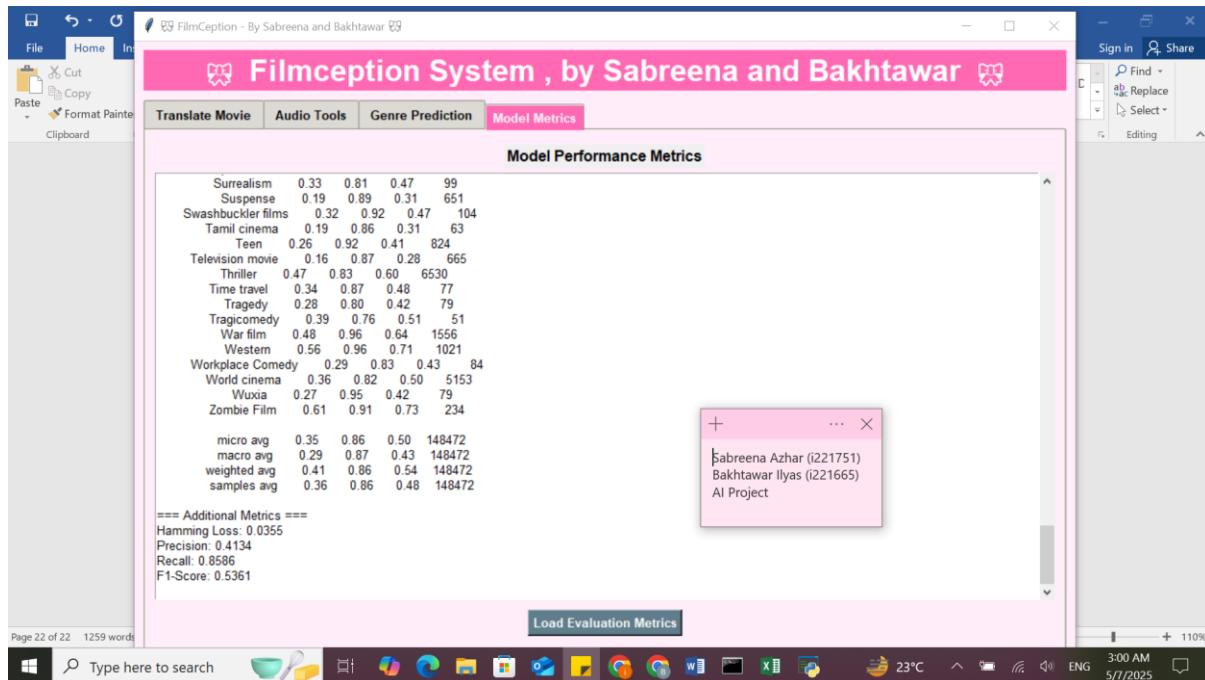
## 4. Model Metrics Tab

### Evaluate Model Performance

- Displays key evaluation metrics:
  - **Hamming Loss**
  - **Precision**
  - **Recall**
  - **F1-Score**



# National University of Computer and Emerging Sciences Islamabad Campus



## • Summary

FilmCeption – Movie Genre Classifier and Translator

**FilmCeption** is an end-to-end machine learning and NLP-based application that processes movie summaries to **predict genres**, **translate content**, and **generate audio**, all wrapped in an interactive user interface. It supports **multi-label classification**, multilingual translation, and audio synthesis, delivering an engaging and accessible user experience.