National University of Computer and Emerging Sciences
Islamabad Campus

Filmception

Static aAn AI-powered Multilingual movie summary translator and genre classifier

Artificial Intelligence

Submitted by: Aleena Fatima, Muhammad Haseeb
Roll number: 22i-2353, 22i-1698
Date: 6 May 2025

Introduction

The Filmception project aims to develop an AI-powered multilingual movie summary translator and genre classifier. This report outlines the steps taken in the project, focusing on data preprocessing and cleaning, text translation and audio conversion, movie genre prediction model development, and the creation of an interactive user interface.

1. Data Preprocessing and Cleaning

Summary Extraction and Cleaning

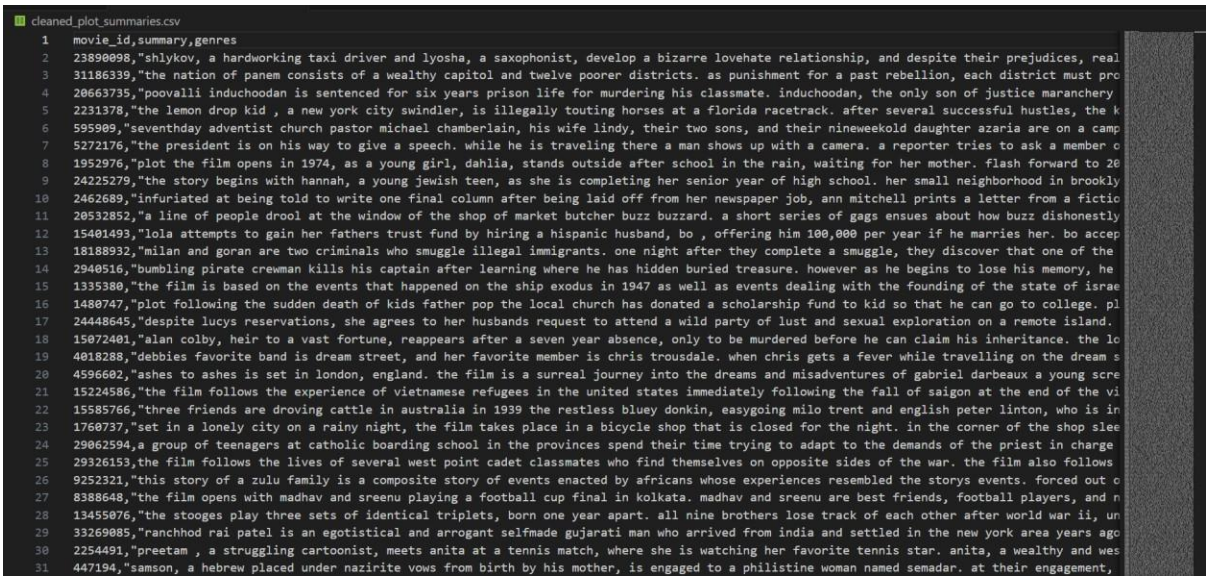
Preprocessing Steps:

- Removing Special Characters: All non-alphanumeric characters were eliminated to ensure clean text.
- Stopword Removal: Commonly used words that do not contribute to the meaning (e.g., "and", "the") were removed.
- Lowercasing: All text was converted to lowercase to standardize the data.
- Tokenization: The summaries were broken down into individual words or phrases for easier analysis.
- Stemming/Lemmatization: Words were reduced to their base or root form to unify different variations of a word.
- Removing Non-Relevant Words: Numbers and punctuation that do not add value to genre classification were removed.

Metadata Extraction

Genre Information: Extracted from the movie.metadata.tsv file, the genre information was identified as multi-labels, allowing each movie to belong to multiple genres (e.g., "Action", "Adventure", "Comedy").

Final Output: A cleaned dataset was created containing the Movie ID, cleaned summary, and genres. This dataset will be used for training the genre prediction model.



cleaned_plot_summaries.csv

```
movie_id,summary,genres
23890098,"shlykov, a hardworking taxi driver and lyosha, a saxophonist, develop a bizarre lovehate relationship, and despite their prejudices, real
31186339,"the nation of panem consists of a wealthy capitol and twelve poorer districts. as punishment for a past rebellion, each district must pro
20663735,"poovalli induchoodan is sentenced for six years prison life for murdering his classmate. induchoodan, the only son of justice maranchery
2231378,"the lemon drop kid , a new york city swindler, is illegally touting horses at a florida racetrack. after several successful hustles, the k
595909,"seventhday adventist church pastor michael chamberlain, his wife lindy, their two sons, and their nineweekold daughter azaria are on a camp
5272176,"the president is on his way to give a speech. while he is traveling there a man shows up with a camera. a reporter tries to ask a member o
1952976,"plot the film opens in 1974, as a young girl, dahlia, stands outside after school in the rain, waiting for her mother. flash forward to 20
24225279,"the story begins with hannah, a young jewish teen, as she is completing her senior year of high school. her small neighborhood in brookly
2462689,"infuriated at being told to write one final column after being laid off from her newspaper job, ann mitchell prints a letter from a fictio
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
15401493,"lola attempts to gain her fathers trust fund by hiring a hispanic husband, bo , offering him 100,000 per year if he marries her. bo accep
18188932,"milan and goran are two criminals who smuggle illegal immigrants. one night after they complete a smuggle, they discover that one of the
2940516,"bumbling pirate crewman kills his captain after learning where he has hidden buried treasure. however as he begins to lose his memory, he
1335380,"the film is based on the events that happened on the ship exodus in 1947 as well as events dealing with the founding of the state of israe
1480747,"plot following the sudden death of kids father pop the local church has donated a scholarship fund to kid so that he can go to college. pl
24448645,"despite lucys reservations, she agrees to her husbands request to attend a wild party of lust and sexual exploration on a remote island.
15072401,"alan colby, heir to a vast fortune, reappears after a seven year absence, only to be murdered before he can claim his inheritance. the lo
4018288,"debbies favorite band is dream street, and her favorite member is chris trousdale. when chris gets a fever while travelling on the dream s
4596602,"ashes to ashes is set in london, england. the film is a surreal journey into the dreams and misadventures of gabriel darbeaux a young scre
15224586,"the film follows the experience of vietnamese refugees in the united states immediately following the fall of saigon at the end of the vi
15585766,"three friends are droving cattle in australia in 1939 the restless bluey donkin, easygoing milo trent and english peter linton, who is in
1760737,"set in a lonely city on a rainy night, the film takes place in a bicycle shop that is closed for the night. in the corner of the shop slee
29062594,a group of teenagers at catholic boarding school in the provinces spend their time trying to adapt to the demands of the priest in charge
29326153,the film follows the lives of several west point cadet classmates who find themselves on opposite sides of the war. the film also follows
9252321,"this story of a zulu family is a composite story of events enacted by africans whose experiences resembled the storys events. forced out o
8388648,"the film opens with madhav and sreenu playing a football cup final in kolkata. madhav and sreenu are best friends, football players, and n
13455076,"the stooges play three sets of identical triplets, born one year apart. all nine brothers lose track of each other after world war ii, un
33269085,"ranchhod rai patel is an egotistical and arrogant selfmade gujarati man who arrived from india and settled in the new york area years ago
2254491,"preetam , a struggling cartoonist, meets anita at a tennis match, where she is watching her favorite tennis star. anita, a wealthy and wes
447194,"samson, a hebrew placed under nazirite vows from birth by his mother, is engaged to a philistine woman named semadar. at their engagement,
```

Train-Test Split: A careful train-test split was performed to prevent data leakage and ensure robust model evaluation. The dataset was divided into training and testing sets, maintaining a balanced representation of genres.

```
Input: The Quick Brown Fox Jumps Over The Lazy Dog! 123
Output: quick brown fox jump lazy dog

Input:
Output:

Input: None
Output:

Input: the the is are
Output: the

Dataset integrity checks passed.
Total: 41796, Train: 33614 (80.42%), Test: 8182 (19.58%)
Sample Train Row:
 movie_id                                                31186339
cleaned_summary    nation panem consists wealthy capitol twelve p...
genres_list        (Action, Action/Adventure, Drama, Science Fict...
Name: 0, dtype: object
Sample Test Row:
 movie_id                                                23890098
cleaned_summary    shlykov hardworking taxi driver lyosha saxopho...
genres_list                                  (Drama, World cinema)
Name: 0, dtype: object
```

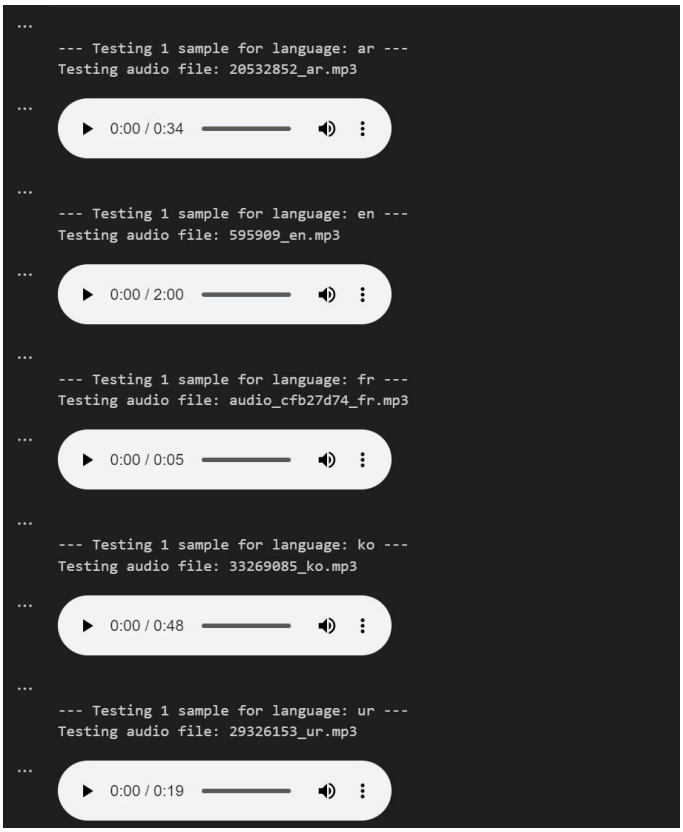
2. Text Translation and Audio Conversion

Text Translation

- Translation Process:
  - The cleaned movie summaries were translated into Arabic, Urdu, and Korean using translation tools such as Google Translate and MarianMT (Hugging Face).
  - A minimum of 50 movie summaries were translated to test the capability to work with multilingual data.

Audio Conversion

- Text-to-Speech (TTS):
  - The translated summaries were converted into audio using TTS engines like gTTS and pyttsx3.
  - The audio files were saved for playback, allowing users to select their preferred language for listening to the movie summaries.



3. Movie Genre Prediction Model

Model Development

- Machine Learning Model:
  - A multi-label classification model was built to predict movie genres based on their summaries.
  - Various model architectures were considered, including Logistic Regression, Random Forests, and more complex models like LSTMs and Transformer-based models (e.g., BERT).

Feature Extraction

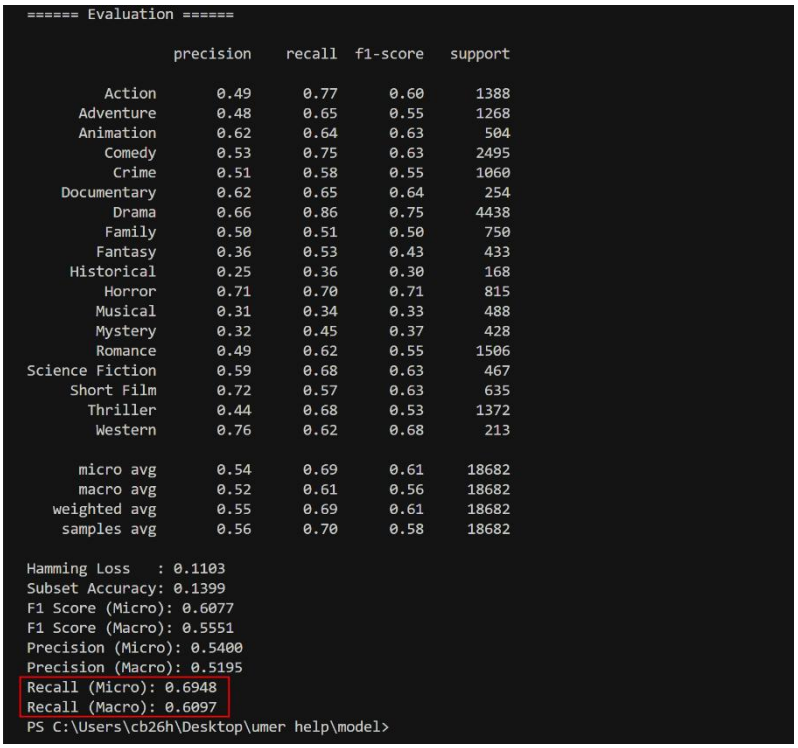
- Techniques Used:
  - Features were extracted from the movie summaries using methods such as TF-IDF and word embeddings (Word2Vec, GloVe).
  - Pre-trained transformers were also utilized for text representation.

Evaluation

- Performance Metrics:
  The model's performance was evaluated using:

  - Accuracy: Overall accuracy of the model.
  - Precision, Recall, F1-Score: Metrics to assess the model's ability to identify specific genres.

- Confusion Matrix: Visualization of classification results for both training and test sets.



```
====== Evaluation ======

                precision    recall  f1-score   support

        Action       0.49      0.77      0.60      1388
     Adventure       0.48      0.65      0.55      1268
     Animation       0.62      0.64      0.63       504
        Comedy       0.53      0.75      0.63      2495
         Crime       0.51      0.58      0.55      1060
   Documentary       0.62      0.65      0.64       254
         Drama       0.66      0.86      0.75      4438
        Family       0.50      0.51      0.50       750
       Fantasy       0.36      0.53      0.43       433
    Historical       0.25      0.36      0.30       168
        Horror       0.71      0.70      0.71       815
       Musical       0.31      0.34      0.33       488
       Mystery       0.32      0.45      0.37       428
       Romance       0.49      0.62      0.55      1506
Science Fiction       0.59      0.68      0.63       467
    Short Film       0.72      0.57      0.63       635
      Thriller       0.44      0.68      0.53      1372
       Western       0.76      0.62      0.68       213

     micro avg       0.54      0.69      0.61     18682
     macro avg       0.52      0.61      0.56     18682
  weighted avg       0.55      0.69      0.61     18682
   samples avg       0.56      0.70      0.58     18682

Hamming Loss   : 0.1103
Subset Accuracy: 0.1399
F1 Score (Micro): 0.6077
F1 Score (Macro): 0.5551
Precision (Micro): 0.5400
Precision (Macro): 0.5195
Recall (Micro): 0.6948
Recall (Macro): 0.6097
PS C:\Users\cb26h\Desktop\umer help\model>
```

4. Interactive User Interface (Menu-Based System)

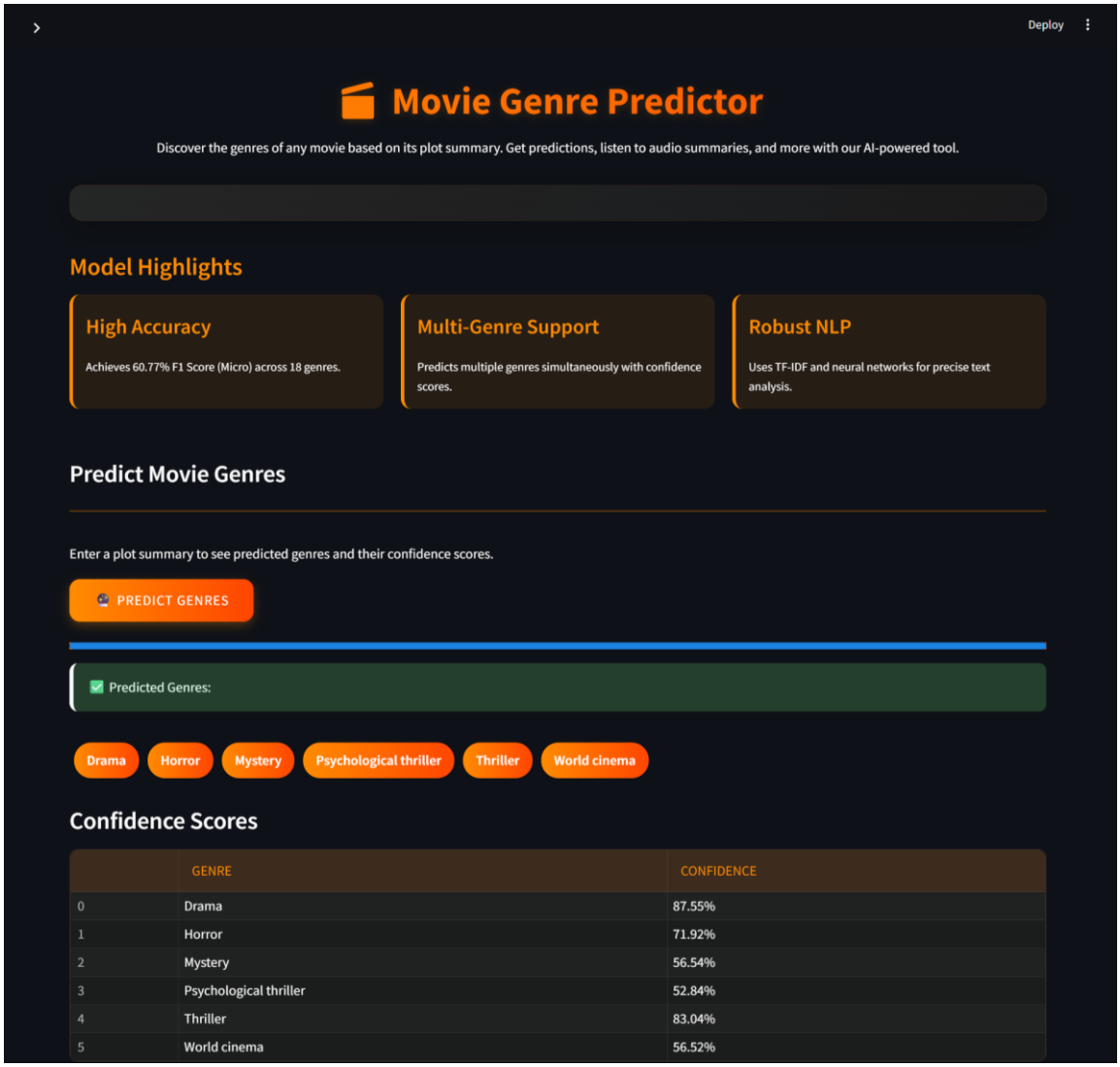
The project features an interactive, menu-driven system designed to enhance user experience. The workflow includes:

Input: Users input a movie summary.

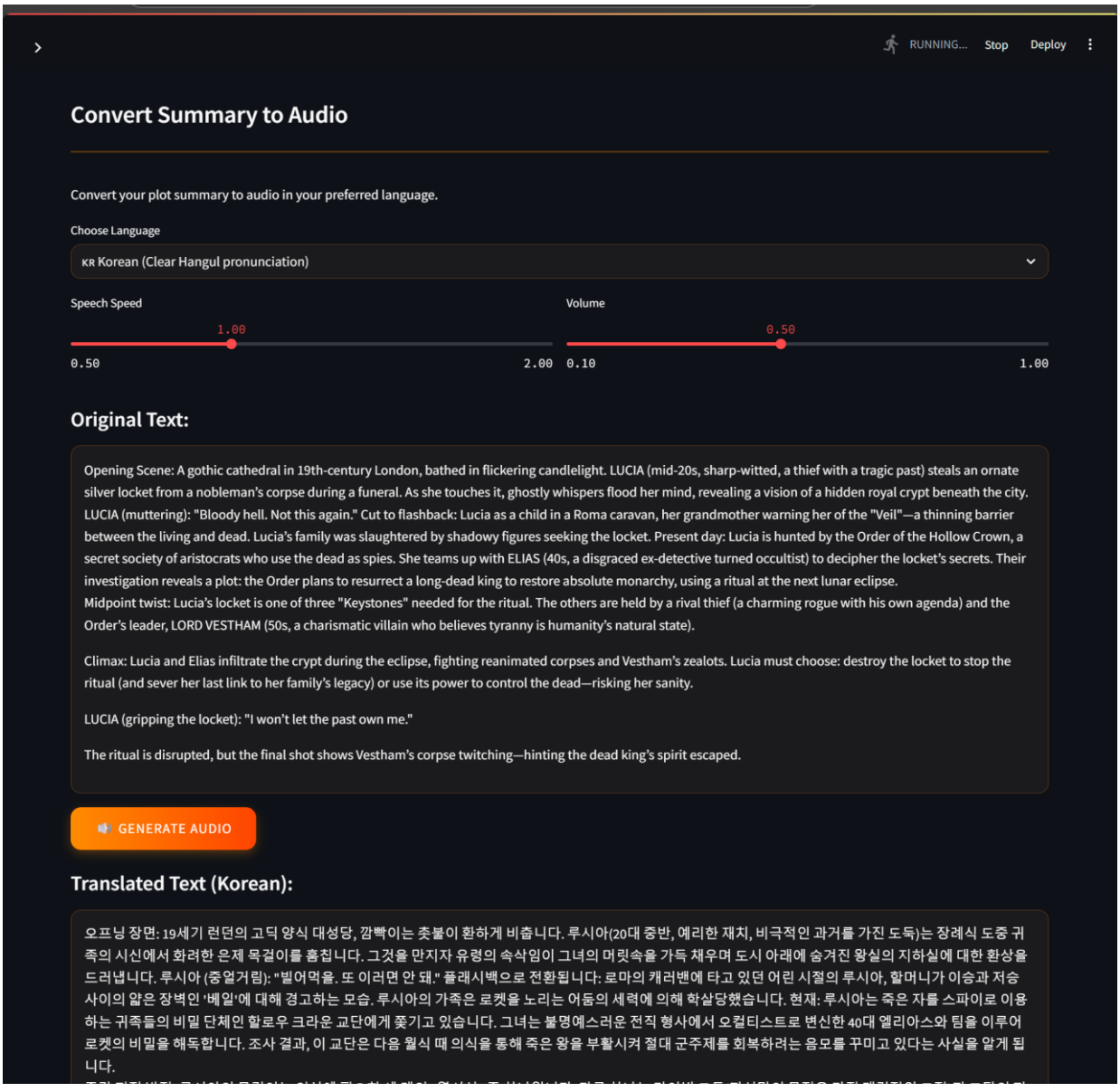
Options:

- Option 1: Convert Summary to Audio ~ Users select the audio language, and the system generates and plays the audio.
- Option 2: Predict Genre ~ Users classify the genre of the movie based on the summary, and the system outputs the predicted genre(s).

The system allows users to explore both features sequentially, enhancing engagement and usability.



- There is translation service with deepl API that gives accurate translation



Conclusion

The Filmception project successfully integrates data preprocessing, multilingual translation, audio conversion, and genre classification into a cohesive system. The interactive user interface further enhances user engagement, making it a comprehensive tool for movie summary translation and genre prediction. Future work may involve refining the model and expanding the dataset for improved accuracy and user experience.